EJSIT
European Journal of Science, Innovation and Technology
ISSN: 2786-4936
www.ejsit-journal.com
Volume 4 | Number 4 | 2024

The Design and Construction of a Big Data System for Water Resource Warning and Forecasting in Vietnam

Nguyen Van Loi, Le Anh Tuan, Tran Duc Thinh, and Dang Tran Trung*
National Center for Water Resources Planning and Investigation, MONRE, Sai Dong ward, Long Bien District, Hanoi, Vietnam (NAWAPI)

ABSTRACT

The study in this paper introduces the importance of a big data system in water resource forecasting, helping to mitigate the damage caused by natural disasters. With increasing demand for observational data and water resource forecasting information, the paper emphasizes the need to improve forecasting technology to provide accurate and timely information. The paper provides an overview of Big Data technology, including large data volumes, fast processing speeds, data diversity, and reliability. The architecture of the Big Data system at the Center for Water Resource Forecasting and Warning is designed with multiple components such as data sources, storage systems, batch processing, and real-time processing. Technologies like Hadoop and HDFS are applied to manage and store distributed data, ensuring data safety and recovery.

The paper also proposes storage infrastructure and data analysis tools such as Hadoop, Apache Spark, along with security measures like data encryption and access control. Finally, the authors emphasize that the application of Big Data in water resource forecasting in Vietnam requires significant investment in infrastructure, technology, and specialized personnel.

Keywords: Big Data, Vietnam Water Resources, Hadoop, HDFS, Distributed Data Storage

INTRODUCTION

Water resource forecasting is crucial for the development of the economy, society, national security, and especially for proactive disaster prevention and mitigation of water-related damage. As society advances, the demand for observational data and information on water resource forecasts continues to grow, not only in terms of content variety but also in the accuracy of warning and forecasting information. According to water resource experts, there are three main reasons why it is essential to evaluate the quality of water resource warnings and forecasts: to monitor forecast quality and determine how accurate the observational data and forecasts are, and whether this accuracy improves over time; to improve the quality of water resource forecasting by identifying what is forecasted incorrectly and how it is wrong to enhance forecasting technology; and to compare the forecast quality between different systems. In this report, the project team will review the collected data, analyze and evaluate it, and identify the management subjects that need to be incorporated into the database.

Overview of Big Data Technology

Big Data technology refers to a collection of technologies, tools, and methods designed to process massive, complex, and unstructured data that traditional data systems cannot handle efficiently. Big Data is characterized by four main aspects: volume, variety, velocity, and veracity.

- Volume: The defining feature of Big Data is its enormous size, which far exceeds the storage capacity of traditional systems. With the increasing amount of data generated daily, traditional storage media such as floppy disks and hard drives are no longer

* Corresponding Author

sufficient. Cloud technologies have become essential for storing and managing this vast amount of data (Mayer-Schönberger, 2013).

- Velocity: Velocity refers to both the speed at which data is generated and the need to process it in real-time. This aspect is particularly important in sectors like finance, telecommunications, and defense, where real-time data processing ensures timely responses to changing conditions. Big Data systems are designed to handle data that needs to be processed immediately upon generation, often within milliseconds (Chen et al., 2014).
- Variety: Unlike traditional data, which is typically structured, over 80% of today's data is unstructured, comprising documents, blogs, images, and other media. Big Data allows for the integration and analysis of these diverse data formats, enabling more comprehensive insights across various sectors (Gandomi & Haider, 2015).
- Veracity: One of the most challenging aspects of Big Data is ensuring the reliability and accuracy of the data being processed. With the rise of social media and mobile interactions, it is becoming increasingly difficult to determine the trustworthiness of data. Thus, filtering out inaccurate or noisy data has become a crucial component of Big Data analytics (Zikopoulos et al., 2012).
- Value: The ultimate purpose of Big Data lies in the value it can provide. Organizations invest in Big Data technologies to extract actionable insights that can drive decision-making and innovation. The value of Big Data is often considered the most important aspect, as it justifies the implementation and scaling of such systems (Laney, 2001).

By addressing these key characteristics—volume, velocity, variety, veracity, and value—Big Data technologies provide organizations with the capability to manage, analyze, and derive insights from increasingly complex datasets.

Big Data Architecture

Big Data architecture refers to the framework that enables the storage, processing, and analysis of large volumes of data. The architecture generally includes several key components:

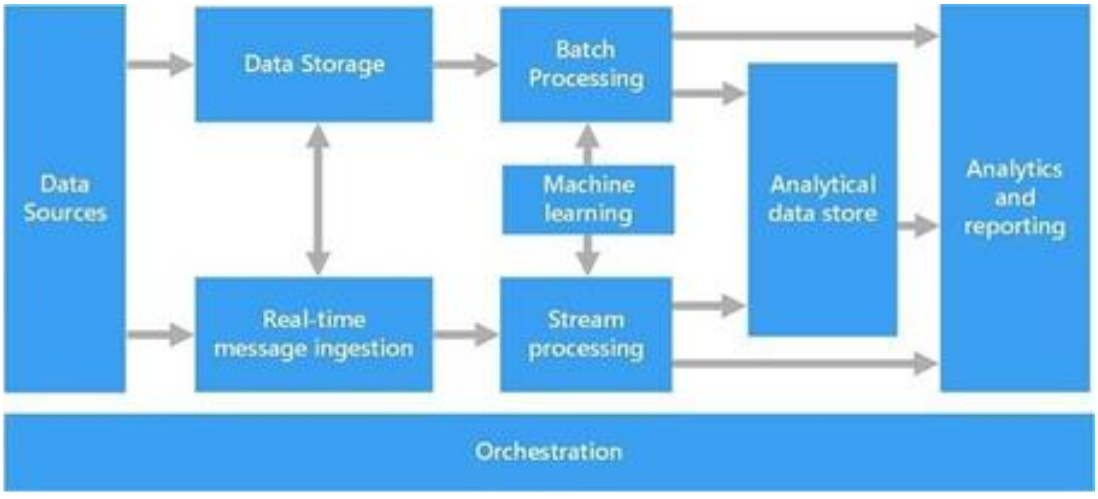

Figure 1. Big data Architecture

Data Sources: This is the origin of data, which can include structured, unstructured, and semi-structured data. Data sources can vary significantly, including data from applications, relational databases (e.g., transactions from retail systems, bank transfers), logs generated by applications (e.g., system processing time logs), or real-time data from IoT devices (e.g., images from cameras, temperature and humidity sensors). Big Data frameworks are designed to ingest and manage this diverse range of data types (Chen et al., 2014).

Data Storage: Data storage is designed to handle vast volumes of data in multiple formats generated by the data sources. In Big Data systems like Hadoop, the most commonly implemented model for data storage is distributed file systems, which store data across multiple nodes in a cluster. This ensures that data is safely replicated and can be processed efficiently. Apache Hadoop's HDFS (Hadoop Distributed File System) is one of the most widely used solutions for implementing this component (Zikopoulos et al., 2012).

Batch Processing: Batch processing allows for the processing of large amounts of data by reading data from source files, filtering data based on certain conditions, performing calculations, and then writing the results to an output file. Popular tools for batch processing include Apache Spark, Hive, and MapReduce, which support multiple programming languages such as Java, Scala, and Python (Dean & Ghemawat, 2008).

Real-Time Message Ingestion: Data generated from sources, such as IoT devices, often requires real-time ingestion and storage. This component allows a Big Data system to capture and store various types of real-time data for streaming processing. Apache Kafka is the most popular tool for real-time message ingestion, with other options like RabbitMQ and ActiveMQ also widely used (Kreps et al., 2011).

Stream Processing: Similar to batch processing, stream processing allows real-time data to be filtered, calculated, and stored. It is essential for real-time analytics and decision-making. Tools like Apache Storm and Spark Streaming are commonly used for stream processing (Neumeyer et al., 2010).

Analytical Data Store: This component stores processed data in structured formats to support data analysis tools (BI tools). Data can be stored in an OLAP model using the Kimball Star Schema or in NoSQL technologies like HBase and Cassandra for non-relational data storage (Kimball & Ross, 2019).

Analysis and Reporting: This layer provides users with self-service access to data, allowing them to visualize, analyze, and generate reports. Tools used for data visualization and reporting vary widely, ranging from open-source solutions like D3.js and Dygraphs to commercial products like Tableau and Power BI. Programming languages like Python are also frequently used for custom data visualization (Few, 2012).

Orchestration: Orchestration ensures that all tasks within the Big Data system run smoothly, from data ingestion to storage, filtering, and processing. Tools like Apache Oozie and Airflow are commonly used for orchestrating workflows in Big Data environments (Islam et al., 2012).

SOLUTIONS FOR BUILDING A BIG DATA MANAGEMENT SYSTEM

Suitable Data Sources for Extraction

The collection of diverse and rich data from multiple sources is fundamental to building a Big Data management system aimed at water resource forecasting and warning. Below are the key data sources:

Data from Sensors and Environmental Monitoring Stations: Information from water level sensors, water quality monitors, and weather data collected at national and local environmental monitoring stations provides critical insights into the environmental and water resource conditions (Fan, McCook, & Yen, 2015).

Satellite Data: Satellite data from sources like Landsat, Sentinel, and MODIS are used to collect information on water surface area, rainfall forecasts, and environmental changes, offering a comprehensive view of water resources (Gao & Liu, 2016).

Simulation and Modeling Data: Data from simulation models are used to predict and identify changes in water resources based on various climate change scenarios, assisting in future planning and forecasting (Ahmed et al., 2017).

Social Data: Social data provide information on water consumption, population, and industrial activity, helping to understand the human impact on water resources (Gleick, 2014). This type of data can be critical for evaluating water demand and supply patterns.

Historical and Statistical Data: Historical data on water levels, rainfall, and water source variations allow for trend evaluation and comparison with current data to provide deeper insight into water resource dynamics (Vorosmarty et al., 2000).

Ecological and Biodiversity Data: This data offers information on the impact of environmental changes on water resources, including biodiversity and ecosystem health, which is crucial for sustainable water resource management (Zedler & Kercher, 2005).

Cross-Sector and Open Data: Integrating information from various sectors such as agriculture, healthcare, and economics allows for a more comprehensive assessment of water resource conditions. Open data initiatives can further enhance transparency and collaboration in water management (Kitchin, 2014).

By combining and leveraging these diverse data sources, a Big Data management system can provide accurate and comprehensive information to support timely and effective water resource forecasting and warnings, helping to mitigate risks and ensure sustainable management of water resources.

Storage Infrastructure

Storing large-scale data requires robust and flexible infrastructure to ensure data integrity, scalability, and easy access. Below are key components of the storage infrastructure:

Cloud Storage Systems: Cloud storage services like Amazon S3, Google Cloud Storage, or Microsoft Azure Storage are widely used for data storage. These services provide flexible scalability, high security, and convenient remote access, allowing organizations to manage large volumes of data efficiently (Marston et al., 2011).

Distributed Storage Systems: Distributed storage infrastructure is built using multiple storage servers and physical storage devices at different locations. This ensures data integrity and recovery in case of system failures, which is crucial for maintaining operational resilience in Big Data environments (Shvachko et al., 2010).

NoSQL Data Storage: NoSQL databases, such as MongoDB, Cassandra, or HBase, are well-suited for storing unstructured data. These databases allow for flexible data models and are optimized for managing large-scale, diverse datasets from various sources (Cattell, 2011).

Backup and Recovery Systems: Implementing periodic backup systems ensures data safety. Clear and well-documented processes for data backup and recovery are essential to maintaining data integrity. Regular testing of backup and recovery procedures is crucial for ensuring that systems function as expected during an emergency (Li et al., 2009).

Data Security Integration: Security measures such as data encryption, access control, and logging are essential to protecting data from external and internal threats. By integrating strong security protocols, organizations can safeguard sensitive information and ensure compliance with data protection regulations (Subashini & Kavitha, 2011).

Metadata Integration and Data Management: Using metadata to describe and manage data enhances the efficiency of data search and retrieval. Proper metadata management is key to understanding the structure and value of data, enabling better data governance and utilization (Vermesan & Friess, 2013).

Continuous Evaluation and Infrastructure Optimization: Regular assessments and optimization of the storage infrastructure are necessary to ensure scalability, performance, and flexibility in handling Big Data. This involves monitoring system performance and making adjustments as data volumes and requirements grow (Hashem et al., 2015).

A well-constructed and maintained Big Data storage infrastructure is essential for ensuring data integrity, security, and efficient retrieval. By combining cloud storage,

distributed systems, and NoSQL databases with robust security and management practices, organizations can effectively manage their Big Data environments.

Data Management and Security

Effective data management, combined with robust security measures, is critical in building a Big Data management system, especially when handling sensitive and diverse information from multiple sources. Below are the key aspects of data management and security:

- Metadata Management: Establishing a metadata management system is essential for describing, classifying, and organizing data. Metadata helps in understanding the structure and content of the data, providing crucial information for efficient search and data utilization. Well-managed metadata allows organizations to optimize data governance and accessibility (Katal et al., 2013).
- Data Management Policies: Clear policies must be defined for data storage, processing, and classification based on legal requirements, industry standards, and organizational needs. These policies ensure compliance and promote the effective management of data. Data governance policies guide organizations in structuring, storing, and safeguarding data to meet regulatory demands (Cheong & Chang, 2007).
- Access Control Management: A stringent access control system should be built to monitor who can access specific data and to what extent. Authentication and authorization mechanisms, such as role-based access control (RBAC), help protect data from unauthorized access, ensuring that only those with the right permissions can interact with sensitive information (Ferraiolo & Kuhn, 1992).
- Data Encryption: Applying data encryption is critical to protecting data during transmission, storage, and processing. Strong encryption methods, such as the Advanced Encryption Standard (AES), ensure that data cannot be accessed or read without the correct decryption key. This is particularly important when dealing with sensitive or personal data (Stallings, 2006).
- High-Level Security Controls: High-level security measures, such as proxy servers, Virtual Private Networks (VPNs), and firewalls, should be employed to safeguard the system from cyberattacks and unauthorized access. These controls help in shielding the network and data from external threats (Zissis & Lekkas, 2012).
- Sensitive Data Protection: For highly sensitive data, additional protective measures must be implemented, including continuous monitoring, advanced encryption, and restricted access based on the "need-to-know" principle. These measures are essential to ensure the security and confidentiality of critical data (Elmagarmid & Bertino, 2005).
- Security Awareness and Training: Regular training sessions should be organized to enhance employees' awareness of cybersecurity and information protection. Such training helps staff better understand risks and the necessary measures to safeguard data, promoting a culture of security within the organization (Furnell & Vasileiou, 2017).

Proper data management, along with suitable security measures, ensures the integrity, confidentiality, and compliance of Big Data within the management system. By establishing clear policies, implementing advanced security technologies, and fostering security awareness, organizations can effectively manage and secure their Big Data environments.

Data Analysis and Exploitation Tools

Analyzing Big Data requires the use of powerful and flexible tools to extract valuable insights. Below are some of the key Big Data analytics tools and a simple explanation of how they operate:

Apache Hadoop: This is an open-source framework used for processing and storing large-scale data. Hadoop utilizes the MapReduce model to analyze data by breaking it down into smaller pieces and processing them in parallel across multiple nodes, enabling efficient data handling and computation at scale (Dean & Ghemawat, 2008).

Apache Spark: Spark is a robust Big Data processing tool with real-time data handling capabilities. It provides libraries for performing complex analyses and accelerates data processing by utilizing in-memory computing. This makes Spark significantly faster than traditional disk-based frameworks like Hadoop MapReduce (Zaharia et al., 2010).

Python and R: These are widely used programming languages in Big Data analytics. Both languages offer powerful libraries for statistical analysis, data processing, and data visualization. Python is particularly known for its ease of use and flexibility, while R is favored for its advanced statistical capabilities (Muenchen, 2011)

SQL and NoSQL Databases: SQL (Structured Query Language) is used to query and analyze data in relational databases, while NoSQL databases like MongoDB and Cassandra are employed to handle unstructured or complex structured data. NoSQL solutions are essential for managing diverse data types that don't fit into the rigid structure of traditional relational databases (Cattell, 2011).

Data Visualization Tools: Tools such as Tableau, Power BI, and QlikView are used to visualize Big Data through charts, graphs, and maps. These tools help users understand data more clearly through interactive visualizations, facilitating data-driven decision-making (Few, 2012).

Machine Learning and AI Technologies: Machine learning and artificial intelligence (AI) are used for predictive analytics, anomaly detection, and optimization in Big Data processing. These technologies automate the discovery of patterns in data and predict future trends, providing significant value in areas like fraud detection and demand forecasting (Jordan & Mitchell, 2015).

These Big Data analytics tools allow organizations to efficiently extract meaningful insights from vast datasets, from initial data processing to advanced visualization and deep analysis. Leveraging these tools helps businesses unlock the full potential of their data.

APPLICATION OF BIG DATA TECHNOLOGY

Big Data File Storage Technology Using Hadoop Distributed File System (HDFS)

Hadoop Distributed File System (HDFS) is a distributed data storage system developed by Hadoop. It divides data into smaller blocks and stores them across different nodes within a cluster. HDFS offers high availability and scalability, ensuring efficient access to data stored across the Hadoop clusters.

HDFS is implemented on moderately priced hardware, as server hardware can often experience failures. For this reason, HDFS is designed with high fault tolerance to minimize risks and reduce downtime. HDFS fragments large data into smaller pieces, distributing these across different nodes, and replicates the smaller pieces across several nodes. Therefore, when one node experiences a failure, the system automatically retrieves the data from another node, ensuring seamless processing continuity. This fault-tolerance capability is a key feature of HDFS (Shvachko et al., 2010).

- HDFS Architecture

HDFS follows a master/slave architecture. An HDFS cluster always consists of one NameNode, which acts as the master server, responsible for managing the file system and coordinating file access. In HDFS, large files are divided into one or more blocks, which are stored on a set of DataNodes.

The primary tasks of the NameNode include opening, closing, and renaming files and directories. It also regulates access to the file system, while the DataNodes are responsible for reading and writing to the file system. Additionally, DataNodes handle creating, deleting, and replicating data based on instructions from the NameNode. The NameNode and DataNode work as follows:

- NameNode: This component coordinates client interactions with the HDFS system. Since the DataNodes store the actual data blocks of files in HDFS, they respond to the access requests. The NameNode carries out its tasks through a daemon that runs on port 8021 (Shvachko et al., 2010).
- DataNode: Each DataNode server runs a datanode daemon on port 8022. Periodically, each DataNode reports to the NameNode the list of all blocks it is storing, allowing the NameNode to update its metadata accordingly. After each update, the metadata on the NameNode remains consistent with the data on the DataNodes. This consistent state of metadata is referred to as a checkpoint. Each checkpointed metadata can be used to replicate metadata and restore the NameNode if it fails (Borthakur, 2007).

o Advantages of HDFS:

Data Distribution: In a Hadoop cluster with, for example, 20 machines, a single file can be automatically divided into many parts and stored across the 20 machines. This distribution ensures efficient data processing.

Parallel and Distributed Processing: HDFS allows tasks to be processed in parallel across multiple machines, which significantly reduces processing time compared to using a single machine.

File Replication: This feature helps prevent data loss. If one machine in the cluster experiences a failure, the data is replicated on another machine, ensuring that the system continues to function without losing data.

Vertical Scaling: HDFS allows for system upgrades by increasing the capacity of the existing machines, a feature known as vertical scaling or "scale-up."

Horizontal Scaling: This feature enables the system to be expanded by adding more machines to the cluster, rather than upgrading the hardware of existing machines (Ghemawat et al., 2003).

o Disadvantages of HDFS

Complex Deployment and Management: Deploying and managing a Hadoop cluster requires specialized knowledge of systems, networking, and databases. This can be a barrier for organizations or individuals lacking experience in these areas (Zaharia et al., 2012).

Inconsistent Performance: Performance in a Hadoop cluster can be inconsistent, particularly when data is unevenly distributed among nodes, or when some nodes perform slower than others. The MapReduce framework, a core component of Hadoop, can introduce significant overhead. Tasks that are small or not well-suited to the MapReduce model may result in poor performance.

Limited Scalability: Although Hadoop can scale horizontally (scale-out), expanding the cluster is not always straightforward. Scalability issues may arise in managing and maintaining performance as the cluster grows (Dean & Ghemawat, 2008).

High Resource Requirements: Hadoop requires substantial resources, from hardware to network bandwidth, to operate efficiently (White, 2012).

Difficulty Integrating with Existing Systems: Integrating Hadoop with existing systems can be challenging due to the complexity of pre-built systems or incompatibility with other technologies (Jagadish et al., 2014).

Big Data System Design at NAWAPI for Water Resource Forecasting and Warning

NAWAPI Network Architecture

a) Network Architecture Diagram

The diagram illustrates a network architecture for a Big Data system, specifically focusing on water resource forecasting and warning at NAWAPI (Figure 2).

The system begins with internet signals provided by two Internet Service Providers (VNPT and FPT). These signals pass through a load balancer (Vigor3910) to distribute traffic efficiently across the network.

The architecture includes two layers of firewall (ETX-FW) for security. The first firewall protects the VLAN, which contains various devices like laptops, smartphones, and routers connected through a local area network. The second firewall secures access to the core services, including server management and web services.

The core switch (Core-SW) manages network traffic between the firewalls, connecting the VLAN and web services to the server infrastructure. The server infrastructure is responsible for managing virtual servers and data storage. This setup ensures that network traffic is distributed securely while maintaining high availability and performance for data processing, server management, and storage activities. The system also allows for scaling and redundancy through secure, virtualized resources.

NAWAPI's current storage infrastructure consists primarily of hard drives that are integrated into the server system. In addition, there are two Network-Attached Storage (NAS) devices with a combined total capacity of 85TB. These storage devices are crucial for handling large-scale data collected from various sources for water resource forecasting and warning purposes.

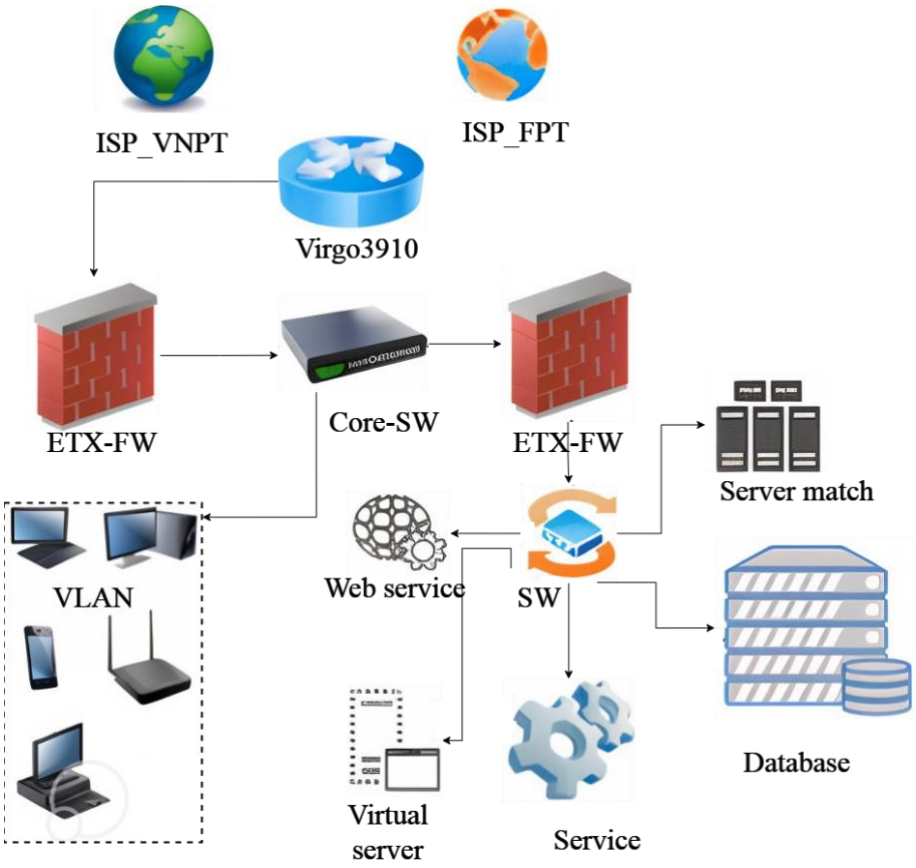

Figure 2. Network Architecture Diagram

b) Technology Infrastructure Architecture

The network system at NAWAPI is provided by two Internet Service Providers (VNPT and FPT), which connect to a load balancing device (Draytek Vigor). Here, the network is routed through a firewall before connecting to a switch, which then connects to various groups of users (Figure 3), including:

- A group of personal computers used internally.
- A group of servers located in the server room. This server system passes through an additional firewall layer before connecting to the network.

At the Draytek load balancer, there is a dedicated network path through a separate switch for the Water Resource Forecasting and Warning room (CEWACO).

The regional divisions connect their data with NAWAPI using QGIS and PgAdmin software to store and manage data efficiently.

This architecture ensures that the network is securely managed through multiple layers of firewalls and switches, while also allowing specialized data flows to dedicated departments like CEWACO for critical water resource forecasting operations. Additionally, the use of software like QGIS and PgAdmin enables efficient data storage and retrieval across connected regional divisions.

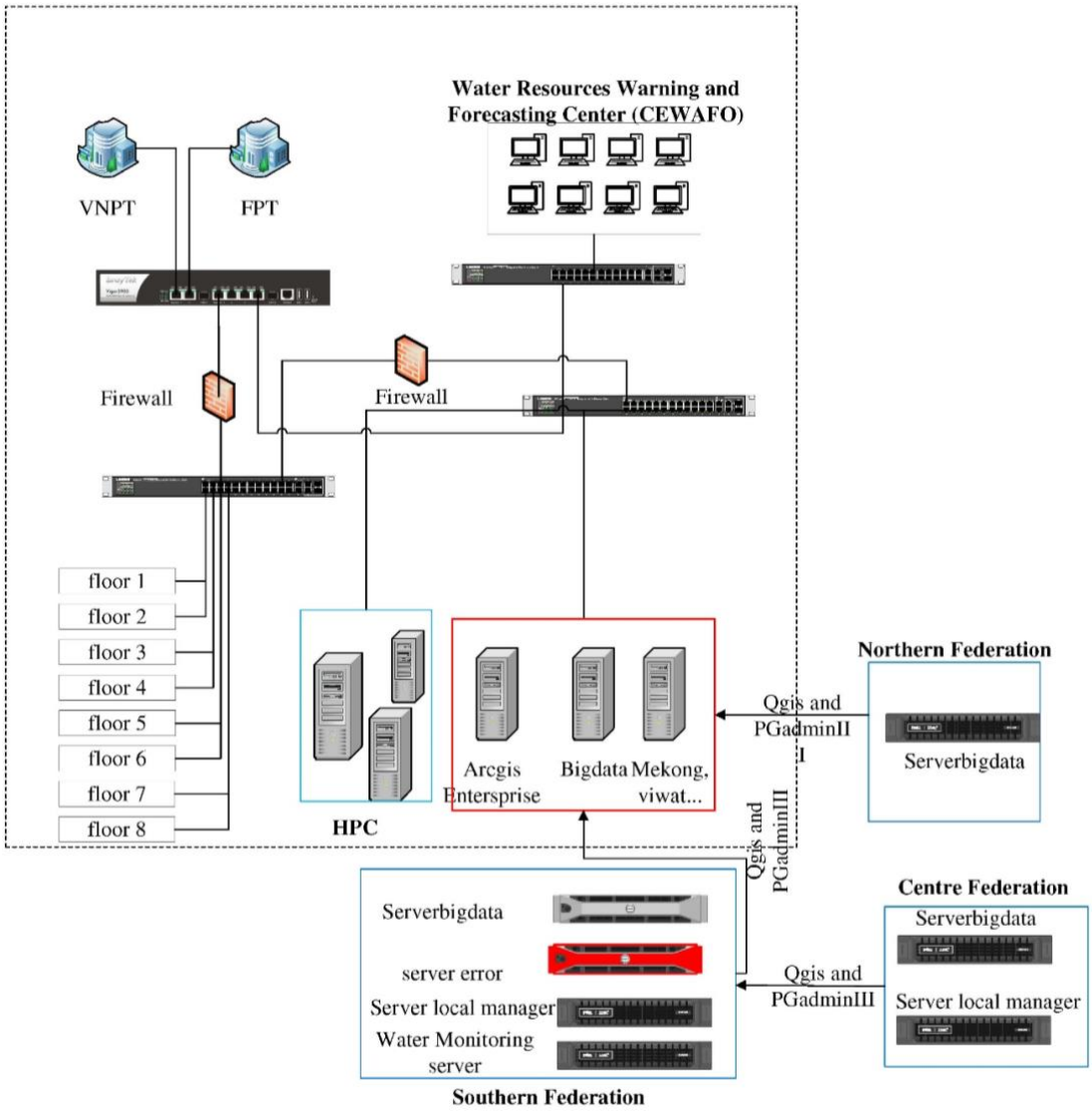

Figure 3. Technology Infrastructure Architecture

BIGDATA System Architecture Components

BIGDATA System Architecture Components of NAWAPI include (Figure 4):

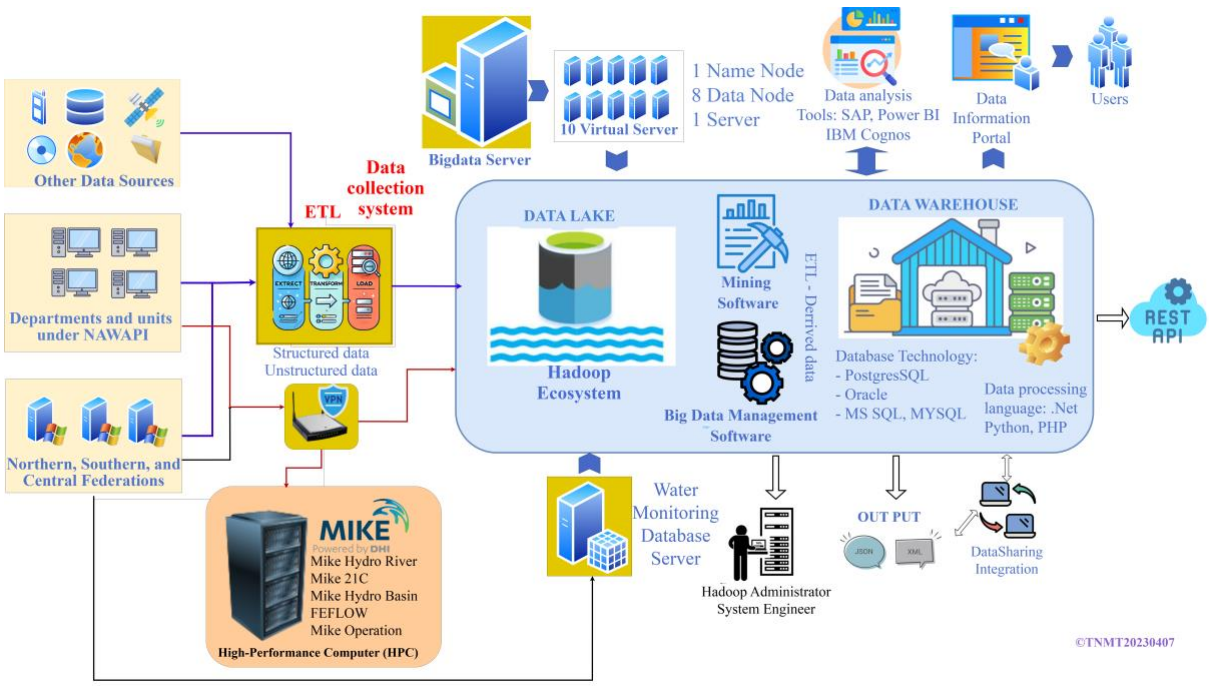

Figure 4. Infrastructure architecture diagram and Water resource big data

a) The components of a Big Data system architecture include:

- Other Data Sources: These are data sources gathered into NAWAPI's Big Data storage system from entities under the management of the Ministry of Natural Resources and Environment, or from other ministries such as the Ministry of Agriculture and Rural Development, and the Ministry of Industry and Trade.
- NAWAPI Departments: These are units that use and provide data directly through NAWAPI's LAN network to the Big Data storage system.
- Northern, Central, and Southern Federations: These are subordinate units of NAWAPI that use and provide data to the Big Data system through NAWAPI's VPN network.
- Big Data Servers: This refers to the physical servers set up for installing, operating, and managing NAWAPI's Big Data system.
- Data Collection System: This is a system of software and tools developed for collecting data from various sources and feeding it into the Big Data system.
- NAWAPI VPN: NAWAPI’s virtual private network, built to ensure secure access to the NAWAPI server system.
- MIKE: This is a water resource management modeling system, serving as both an input data source and an output data provider for NAWAPI’s Big Data system.
- 9 Virtual Servers: These are virtual servers established on physical servers to create datanodes and namenodes for the Big Data system.
- Data Lake and Data Warehouse: The data storage systems for NAWAPI's Big Data operations.
- NAWAPI Database: Databases that NAWAPI has developed in previous stages of the project.
- Specialized Data Exploitation Software: Tools developed for extracting and utilizing water resource data from the Big Data system.

- Data Analysis Tools: SAP BW, IBM Cognos, and Power BI are market-available tools and applications used for data analysis and exploitation from Big Data.

- Information Portal: This is where data is shared and published for organizations and individuals who require it, or for promoting the results of Big Data analysis to the public.

This architecture ensures an integrated and secure flow of data from multiple sources to support NAWAPI's water resource management and forecasting efforts, leveraging advanced tools for analysis and dissemination.

b) Big Data Storage System Architecture at NAWAPI

The architecture of NAWAPI's Big Data storage system consists of one NameNode (NAWAPI NameNode) and multiple DataNodes. Both the NameNode and DataNodes are installed on physical and virtual servers.

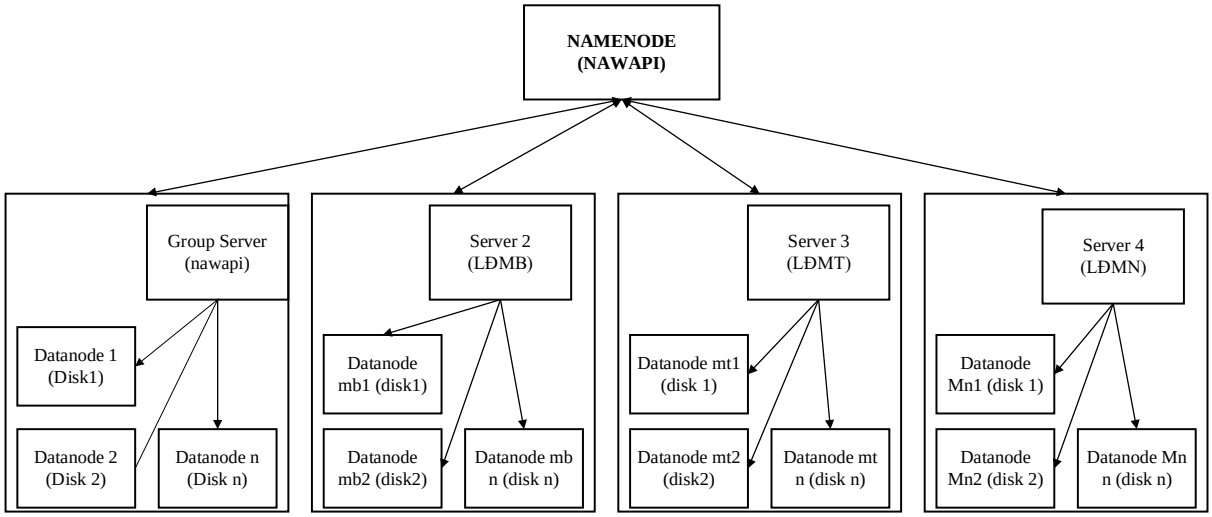

Figure 5. BIGDATA System Architecture Components

- NameNode: The NameNode is the most critical part of the Hadoop Distributed File System (HDFS) architecture at NAWAPI. It is responsible for managing the metadata of files and directories on the distributed file system. Specifically, the NameNode stores information about the directory structure, access permissions, and the physical location of data blocks across the DataNodes. When an application or user requests access to the data, they communicate with the NameNode to locate the specific data blocks (Borthakur, 2007).
- Group Server: This is the physical server infrastructure at NAWAPI, connected via the LAN network. These servers are configured as virtual servers that form the Big Data storage system. Among the virtual servers, one serves as the NameNode, controlling the entire system, while the remaining virtual servers act as DataNodes. Each virtual server is installed on a hard drive with a capacity of 1.2TB.
- Servers at Regional Federations: Each federation under NAWAPI has one server assigned for Big Data system installation. Each server is named and configured for specific tasks as follows:
  - Server 2: This server is used to install a DataNode at the Northern Water Resource Planning and Investigation Federation. Each hard drive on the server is configured as a virtual machine that hosts one DataNode.
  - Server 3: This server is used to install a DataNode at the Central Water Resource Planning and Investigation Federation. Similarly, each hard drive is configured as a virtual machine that hosts one DataNode.

  - Server 4: This server is used to install a DataNode at the Southern Water Resource Planning and Investigation Federation. Each hard drive is configured as a virtual machine that hosts one DataNode.
- DataNode**:** The DataNode is another vital component of NAWAPI's HDFS architecture. Each DataNode is installed on a virtual server and is responsible for storing the actual data in the Big Data system. The DataNodes work in tandem with the NameNode to retrieve, store, and replicate data as needed to ensure the system's fault tolerance and availability (Shvachko et al., 2010).

This distributed architecture ensures that NAWAPI's Big Data system is scalable, fault-tolerant, and capable of handling large volumes of data across multiple regions efficiently.

CONCLUSION

The research and implementation of a Big Data system for water resource forecasting and warning in Vietnam, specifically at NAWAPI, highlight the importance of leveraging modern technologies to address critical challenges. Through the integration of advanced tools like Hadoop Distributed File System (HDFS), Apache Spark, and comprehensive data management systems, NAWAPI has been able to construct a robust infrastructure capable of handling large-scale data across multiple sources.

This system architecture is designed with scalability and fault tolerance at its core, ensuring that the vast amounts of data generated by environmental monitoring stations, satellite imagery, and water resource models can be effectively managed and processed in real time. By utilizing a distributed storage system across various servers, including virtual machines acting as DataNodes, NAWAPI ensures high availability, data redundancy, and performance optimization.

Furthermore, the integration of security measures such as data encryption, access control, and secure VPN connections has strengthened the reliability of the system, safeguarding sensitive data from external and internal threats. Tools like QGIS and PgAdmin further enable efficient data sharing and analysis, fostering collaboration between different regions and sectors involved in water resource management.

In conclusion, NAWAPI’s Big Data system serves as a critical tool in the proactive management and forecasting of water resources, helping mitigate the risks associated with natural disasters and improving decision-making processes in water resource management. Continuous investment in infrastructure, technology, and expertise will be essential for maintaining and expanding this system’s capabilities in the future.

ACKNOWLEDGMENT

The authors would like to express their gratitude to the project “Research and Development of Big Data Analytics Software to Support Water Resource Forecasting and Warning.” Project code: TNMT.2023.04.07 of the Ministry of Natural Resources and Environment (MONRE) and the "Improvement of Groundwater Protection in Vietnam (IGPVN)" project for their assistance and support in this research.

REFERENCES

Ahmed, S., Kim, D., & Kang, C. (2017). Simulation Modeling of Water Resources Under Climate Change Scenarios. *Water Resources Management*, *31*(12), 3769-3785.

Borthakur, D. (2007). The hadoop distributed file system: Architecture and design. *Hadoop Project Website*.

Cattell, R. (2011). Scalable SQL and NoSQL data stores. *Acm Sigmod Record*, *39*(4), 12-27.

Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. *Mobile networks and applications*, *19*, 171-209.

Cheong, L. K., & Chang, V. (2007). The need for data governance: a case study. *ACIS 2007 proceedings*, 100.

Dean, J., & Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. *Communications of the ACM*, *51*(1), 107-113.

Elmagarmid, A. K., & Bertino, E. (2005). *Security in Distributed, Grid, and Pervasive Computing*. CRC Press.

Fan, J., McCook, A., & Yen, S. (2015). Sensor Networks for Environmental Monitoring and Water Resource Management. *International Journal of Sensor Networks*, *17*(2), 134-145.

Ferraiolo, D. F., & Kuhn, D. R. (1992). *Role-Based Access Controls*. 15th National Computer Security Conference.

Few, S. (2012). *Show Me the Numbers: Designing Tables and Graphs to Enlighten*. Analytics Press.

Furnell, S., & Vasileiou, I. (2017). Security education and awareness: just let them burn? *Network Security*, *2017*(12), 5-9.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International journal of information management*, *35*(2), 137-144.

Gao, Y., & Liu, J. (2016). Satellite Remote Sensing for Water Resources Monitoring: A Review. *Journal of Hydrology*, *540*, 408-425.

Ghemawat, S., Gobioff, H., & Leung, S. T. (2003, October). The Google file system. In *Proceedings of the nineteenth ACM symposium on Operating systems principles* (pp. 29-43).

Gleick, P. H. (2014). Water, drought, climate change, and conflict in Syria. *Weather, climate, and society*, *6*(3), 331-340.

Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khan, S. U. (2015). The rise of "big data" on cloud computing: Review and open research issues. *Information systems*, *47*, 98-115.

Islam, M., Huang, A. K., Battisha, M., Chiang, M., Srinivasan, S., Peters, C., ... & Abdelnur, A. (2012, May). Oozie: towards a scalable workflow management system for hadoop. In *Proceedings of the 1st ACM SIGMOD workshop on scalable workflow execution engines and technologies* (pp. 1-10).

Jagadish, H. V., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J. M., Ramakrishnan, R., & Shahabi, C. (2014). Big data and its technical challenges. *Communications of the ACM*, *57*(7), 86-94.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255-260.

Katal, A., Wazid, M., & Goudar, R. H. (2013, August). Big data: issues, challenges, tools and good practices. In *2013 Sixth international conference on contemporary computing (IC3)* (pp. 404-409). IEEE.

Kimball, R., & Ross, M. (2019). *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*. Wiley.

Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures and their consequences*. Sage.

Kreps, J., Narkhede, N., & Rao, J. (2011, June). Kafka: A distributed messaging system for log processing. In *Proceedings of the NetDB* (Vol. 11, No. 2011, pp. 1-7).

Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. *META group research note*, *6*(70), 1.

Li, X., Li, Y., Liu, T., Qiu, J., & Wang, F. (2009, September). The method and tool of cost analysis for cloud computing. In *2009 IEEE International Conference on Cloud Computing* (pp. 93-100). IEEE.

Marston, S., Li, Z., Bandyopadhyay, S., Zhang, J., & Ghalsasi, A. (2011). Cloud computing—The business perspective. *Decision support systems*, *51*(1), 176-189.

Mayer-Schönberger, V. (2013). *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.

Muenchen, R. A. (2011). *R for SAS and SPSS users*. Springer Science & Business Media.

Neumeyer, L., Robbins, B., Nair, A., & Kesari, A. (2010, December). S4: Distributed stream computing platform. In *2010 IEEE International Conference on Data Mining Workshops* (pp. 170-177). IEEE.

Shvachko, K., Kuang, H., Radia, S., & Chansler, R. (2010, May). The hadoop distributed file system. In *2010 IEEE 26th symposium on mass storage systems and technologies (MSST)* (pp. 1-10). IEEE.

Stallings, W. (2006). *Cryptography and network security, 4/E*. Pearson Education India.

Subashini, S., & Kavitha, V. (2011). A survey on security issues in service delivery models of cloud computing. *Journal of network and computer applications*, *34*(1), 1-11.

Vermesan, O., & Friess, P. (2013). *Internet of things: converging technologies for smart environments and integrated ecosystems*. River publishers.

Vorosmarty, C. J., Green, P., Salisbury, J., & Lammers, R. B. (2000). Global water resources: vulnerability from climate change and population growth. *Science*, *289*(5477), 284-288.

White, T. (2012). *Hadoop: The definitive guide*. O'Reilly Media, Inc.

Zaharia, M., Chowdhury, M., Das, T., Dave, A., Ma, J., McCauly, M., ... & Stoica, I. (2012). Resilient distributed datasets: A {Fault-Tolerant} abstraction for {In-Memory} cluster computing. In *9th USENIX symposium on networked systems design and implementation (NSDI 12)* (pp. 15-28).

Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S., & Stoica, I. (2010). Spark: Cluster computing with working sets. In *2nd USENIX workshop on hot topics in cloud computing (HotCloud 10)*.

Zedler, J. B., & Kercher, S. (2005). Wetland resources: status, trends, ecosystem services, and restorability. *Annu. Rev. Environ. Resour.*, *30*(1), 39-74.

Zikopoulos, P., Eaton, C., deRoos, D., Detusch, T., & Lapis, G. (2012). *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data (IBM.)*. New York: McGraw. In.

Zissis, D., & Lekkas, D. (2012). Addressing cloud computing security issues. *Future Generation computer systems*, *28*(3), 583-592.