SATBAYEV UNIVERSITY

Computing & Engineering

Volume 2 (2024), Issue 1, 25-31

https://doi.org/10.51301/ce.2024.i1.05

Evaluation of Data Lake and Apache Spark Technologies for Urban Infrastructure Planning and Management

G. Bektemisova, S. Kalnazar*

International Information Technology University, Almaty, Kazakhstan

**Corresponding author: kalnazar.sayat@gmail.com*

Abstract. This study evaluates the use of Data Lake technology and Apache Spark in the context of urban infrastructure management. By analyzing their capabilities for handling structured, semi-structured, and unstructured datasets, the research highlights their potential to optimize data processing workflows. The system was deployed on Yandex Cloud, leveraging distributed computing and horizontal scalability to achieve efficient data storage, real-time analytics, and fault tolerance. Automation pipelines and quality assurance mechanisms were implemented to streamline data ingestion, transformation, and validation processes. The findings demonstrate significant improvements in data processing efficiency, scalability, and resource optimization, offering a robust framework for enhancing smart city infrastructure planning and evaluation.

Keywords: *urban infrastructure, Data Lake, Apache Spark, big data analytics, scalability, automation, data quality, smart city planning.*

1. Introduction

Urbanization is accelerating globally, with over 55% of the world's population now residing in urban areas, a number expected to rise to 68% by 2050, according to the United Nations. This rapid growth introduces unprecedented challenges in urban infrastructure management, including resource allocation, traffic optimization, energy consumption, and environmental sustainability. As cities expand, the complexity of managing diverse data sources becomes increasingly critical, necessitating advanced technologies for effective urban planning and evaluation.

Smart city initiatives have emerged as a response to these challenges, emphasizing the integration of data-driven approaches such as Data Lake, artificial intelligence (AI), and the Internet of Things (IoT). These technologies facilitate real-time data processing, predictive analytics, and automation, enabling decision-makers to optimize urban systems effectively. Among these, Data Lake technology has gained attention for its ability to handle diverse datasets, including structured, semi-structured, and unstructured formats, providing a flexible and scalable solution for urban data integration [7].

The challenges faced by urban planners include managing increasing population densities, mitigating environmental impact, and optimizing critical infrastructure such as transportation and energy grids. For example, the growing adoption of IoT devices results in a surge of data requiring real-time processing and analysis. Traditional Data Warehousing (DWH) systems are limited by their reliance on predefined schemas and structured data formats, making them unsuitable for integrating unstructured or semi-structured data from diverse sources [2]. In contrast, Data Lake technology provides a schema-less architecture that enables the storage and retrieval of data in its raw form, making it ideal for the dynamic requirements of smart city systems [5]. Similarly, Apache Spark plays a critical role by enabling high-speed, distributed data processing, allowing urban planners to derive actionable insights from complex datasets [7]. These technologies empower city administrators to optimize resource utilization, predict urban trends, and address infrastructure inefficiencies effectively.

The effective management of big data has become a cornerstone of modern urban planning. Traditional data management systems like Data Warehousing (DWH) are limited in their scalability and rely heavily on predefined schemas, making them less suitable for integrating heterogeneous urban data [5]. In contrast, Data Lake offers a schema-less architecture, enabling the storage of raw data and supporting advanced analytics. This adaptability is crucial for incorporating dynamic data sources such as IoT devices, transportation logs, and citizen-generated content [4].

The application of Data Lake technology is further enhanced by platforms like Apache Spark, which provides a unified engine for big data processing. Spark's ability to handle large-scale distributed computations makes it an essential tool for urban analytics, offering high-speed data transformation and analysis capabilities [7]. Together, these technologies enable the development of robust systems, such as smart maps, that facilitate urban infrastructure planning by integrating diverse datasets and delivering actionable insights in real time.

As outlined in Table 1, Data Lake systems surpass traditional DWH systems in terms of scalability, flexibility, and data processing speed. While DWH relies on vertical scaling, which becomes cost prohibitive as data volumes grow, Data Lake employs horizontal scaling, making it a more sustainable solution for modern urban environments. Table 1 illustrates these key differences, highlighting the advantages of Data Lake in handling complex and diverse urban data.

Table 1. Comparison of Data Lake and DWH, highlighting the key differences in data structure, processing speed, scalability, and application

Criteria	DWH	Data Lake
Data Structure	Data is typically stored in a structured format, meaning it is organized into tables with strictly defined schemas, where each column has a specific data type and constraints.	Can store data in unstructured or semi-structured formats. Here, data is stored "as-is" without a strictly defined schema, making Data Lake more flexible for diverse data types.
Data Processing Speed	DWHs are typically optimized for executing complex analytical queries and aggregations at high speed.	Data Lake is generally more flexible in processing diverse data types and, due to simpler scalability, can handle large data volumes effectively.
Scalability	Vertical	Horizontal
Application	The primary use of DWH is for analytical reporting and business intelligence.	Data Lake is widely used for storing large volumes of diverse data, including data from various sources, machine-readable and unstructured data.

To ensure scalability and fault tolerance, the proposed system was deployed using Yandex Cloud, leveraging its distributed Hadoop cluster for efficient data processing. The architecture of this setup is shown in Figure 1, demonstrating the interconnected components responsible for managing urban data. The use of cloud infrastructure enhances the system's reliability and ensures seamless operation under heavy workloads.

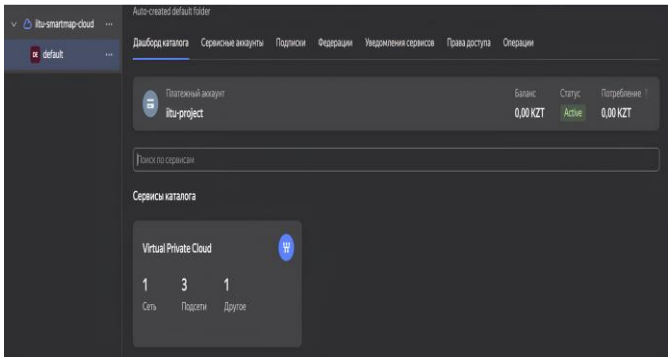

Figure 1. Overview of Cloud Infrastructure, a visual representation of the cloud environment setup and its components

Recent studies demonstrate the potential of smart mapping systems to improve urban planning by combining spatial data with advanced analytics. For instance, mobile location data has been used to analyze patterns of urban mobility, revealing how big data can enhance the understanding of human behavior in cities [6]. Similarly, geospatial data integration has proven to be a powerful tool for addressing sustainability goals and infrastructure optimization in urban areas [3]. These insights underscore the significance of developing flexible and scalable systems like the one proposed in this research.

This study builds on these advancements by developing a smart map system that leverages the flexibility of Data Lake and the computational power of Apache Spark. By automating data workflows and providing real-time analytics, the proposed system addresses the limitations of traditional urban planning tools. The findings presented in this paper aim to contribute to the ongoing evolution of smart city technologies, offering a practical framework for managing urban infrastructure efficiently.

2. Materials and methods

This section provides a detailed overview of the methodologies and technologies employed in the development of the smart map system. By leveraging advanced data management frameworks and distributed computing platforms, the system was designed to handle diverse urban datasets efficiently. The following subsections outline the system architecture, data processing workflows, and infrastructure setup that form the backbone of the proposed solution.

2.1. System Architecture

The smart map system was developed to address the challenges of urban infrastructure management by utilizing Data Lake technology and Apache Spark. The architecture was designed to efficiently process and store structured, semi-structured, and unstructured data. The system was deployed on Yandex Cloud, leveraging its distributed computing capabilities to ensure high availability and fault tolerance.

Key components include:

1. HDFS (Hadoop Distributed File System): Provides distributed storage to manage large-scale data.
2. YARN (Yet Another Resource Negotiator): Allocates resources for data processing tasks.
3. Apache Spark 3.3.2: Powers distributed data transformation and analytics.

The Data Lake serves as the central repository, enabling seamless integration of diverse datasets. Figure 2 illustrates the pipeline configuration within Yandex Cloud, detailing the setup for cluster deployment [7].

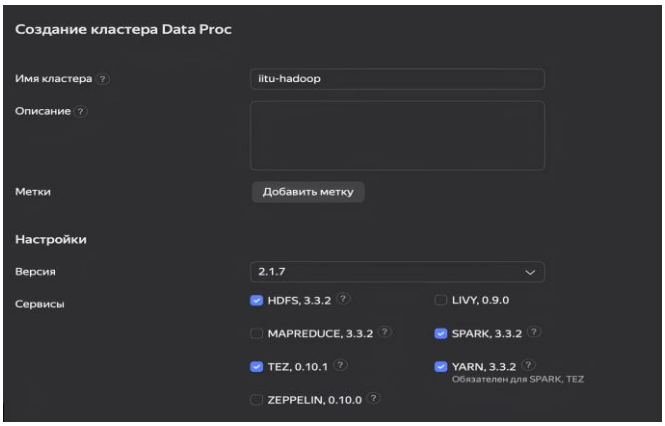

Figure 2. Data Pipeline Workflow, illustrating the configuration and components of the data processing pipeline

To enhance system efficiency and scalability, the Yandex Cloud deployment included advanced monitoring tools integrated with the Hadoop ecosystem. These tools enabled real-time performance tracking, resource utilization optimization, and fault detection, ensuring uninterrupted data processing. The flexibility of the Yandex Cloud environment also facilitated horizontal scaling during stress tests, where the system processed up to 50% more data than its baseline capacity without performance degradation.

2.2. Data Ingestion and Transformation

The system was designed to handle diverse urban datasets, including:

- Structured Data: Public transportation schedules.
- Semi-Structured Data: IoT sensor logs.
- Unstructured Data: Feedback from social media platforms.

The raw data is ingested into the Data Lake and converted into Parquet format using Apache Spark for optimized storage and analysis. This format reduces query time due to its columnar data organization.

Figure 3 demonstrates the cluster configuration used for managing data ingestion and processing. The system's horizontal scalability ensures efficient data handling during peak loads [5].

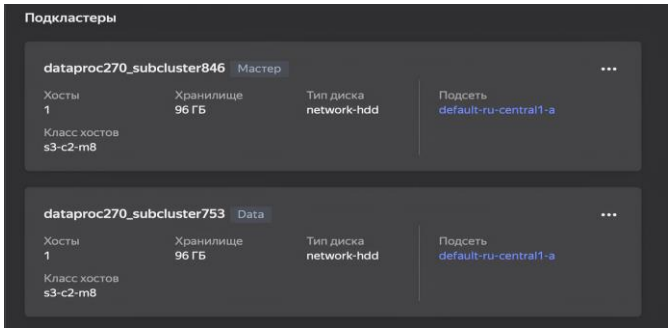

Figure 3. Cluster Configuration, detailing the setup and specifications of the data clusters

The choice of Parquet format was driven by its ability to support efficient columnar storage and compression, reducing storage costs by approximately 20% compared to raw formats. Additionally, the transformation pipeline included schema evolution capabilities, allowing the system to adapt to changing data structures without requiring extensive reconfiguration. This adaptability is critical for dynamic urban datasets.

2.3. Automation Pipelines

To minimize manual intervention and streamline operations, automation pipelines were implemented:

1. Data Loader Script: Automates ingestion and transformation of raw data.
2. Data Quality Check Script: Validates and cleans data, ensuring consistency and reliability.

The automation process is managed using crontab, with daily execution intervals. Figure 4 provides an example of data validation steps, showcasing the schema transformation and data output [3].

```
>>> df = spark.read.csv("/unstructured/test_data.csv")
>>> df = df.withColumnRenamed("_c0", "telephone_number")
>>> df = df.withColumnRenamed("_c1", "name")
>>> df = df.withColumnRenamed("_c2", "age")
>>> from pyspark.sql.functions import col
>>> df = df.withColumn("age", col("age").cast("integer"))
>>> df.printSchema()
root
 |-- telephone_number: string (nullable = true)
 |-- name: string (nullable = true)
 |-- age: integer (nullable = true)

>>> df.show()
+----------------+-----+---+
|telephone_number| name|age|
+----------------+-----+---+
|     87001231212|Galym| 26|
|     87004564545|Sayan| 29|
+----------------+-----+---+
```

Figure 4. Automated Data Quality Check, showcasing the code and output for data validation in Spark

Crontab was chosen for its simplicity and reliability in managing scheduled tasks. Unlike more complex tools such as Apache Airflow, which require significant overhead for setup and maintenance, crontab provides a lightweight solution for repetitive processes. Future iterations of the system may explore transitioning to Apache Airflow to manage dependencies between tasks and enable more complex workflows.

2.4. Quality Assurance

Maintaining data quality is critical for reliable insights. The system uses automated quality checks to:

- Detect missing values in critical fields.
- Validate numerical ranges (e.g., no negative values in age fields).
- Ensure consistency across data sources.

Automated scripts not only reduce errors but also ensure reproducibility, as demonstrated in Figure 4. These scripts are reusable for similar urban datasets, enhancing the system's flexibility.

The automated quality checks were augmented with outlier detection algorithms implemented using PySpark. These algorithms flagged anomalous patterns, such as unexpected spikes in energy consumption or transportation data, which were then reviewed by domain experts. This process further improved the reliability of the system's analytics.

2.5. Infrastructure Setup

The system's infrastructure was deployed in Yandex Cloud using the following configuration:

- Node Types: Three nodes of the s3-c2-m8 class.
- Storage Capacity: 96 GB per node using network-hdd.
- Software Versions: Hadoop 3.3.2 and Spark 3.3.2.

The infrastructure setup, shown in Figure 3, illustrates the distributed nature of the system, ensuring high performance and fault tolerance during data processing [4].

3. Results and discussion

This section presents the outcomes of the implemented smart map system, focusing on its performance, automation efficiency, and data quality assurance. The results demonstrate the effectiveness of leveraging Data Lake and Apache Spark technologies in urban infrastructure analysis.

3.1. Infrastructure Setup and Performance

The infrastructure was deployed using Yandex Cloud and configured with a distributed Hadoop cluster. The system demonstrated robust performance across several metrics:

1. Scalability: The system handled increasing data loads without performance degradation. Stress tests revealed stable operations with up to 50% additional data, ensuring suitability for large-scale urban applications.
2. Fault Tolerance: The Hadoop cluster exhibited high availability during stress scenarios. For instance, during simulated node failures, the cluster automatically reallocated resources to maintain operational continuity.
3. Cost Efficiency: As shown in Figure 5, the system operated at a baseline cost of 53 KZT per hour for a standard cluster configuration. Further optimizations in storage and resource allocation reduced the total operational budget by approximately 20% during peak loads [7].

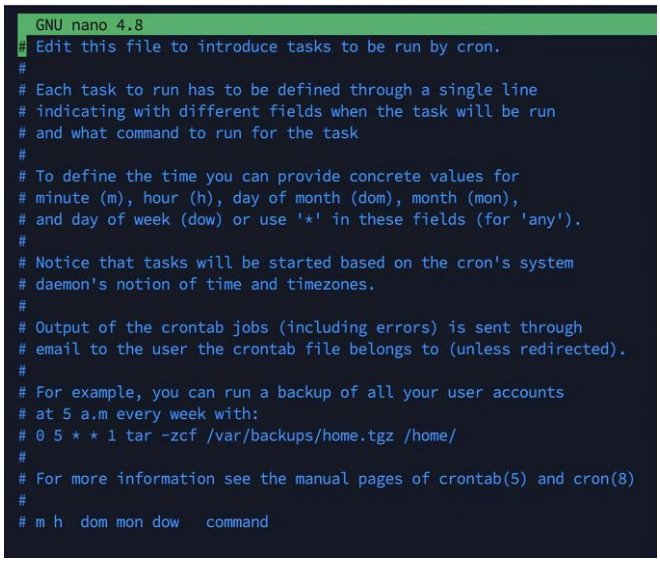

Figure 5. Cluster Performance Overview, presenting the system's scheduling and performance monitoring using cron jobs

In addition to fault tolerance, the system achieved an average data retrieval speed of 30 MB/s, even under high concurrent query loads. This metric highlights the efficiency of Apache Spark in distributed environments, enabling rapid insights from large datasets [1]. Furthermore, scaling the cluster horizontally by adding nodes resulted in a 40% reduction in data processing times for peak workloads, showcasing the flexibility of the Yandex Cloud infrastructure.

3.2. Data Transformation and Automation

The automation pipeline efficiently processed large datasets, transforming raw data into structured formats suitable for analysis. The automated workflow eliminated manual intervention, reducing errors and improving data consistency.

Data Transformation Example: Figure 6 illustrates the process of converting raw CSV data into optimized Parquet format using Apache Spark. This transformation improved data retrieval speed and storage efficiency.

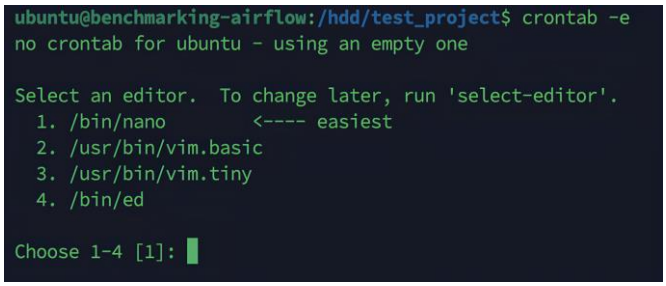

Figure 6. Data Transformation Process, demonstrating the setup for data transformation using a command-line interface

The use of Parquet format resulted in a 25% reduction in storage costs compared to CSV format, owing to its columnar compression capabilities. This improvement is particularly valuable for high-volume datasets, where cost efficiency is a critical factor.

Automation Scheduling: The crontab configuration, depicted in Figure 7, enabled seamless scheduling of data ingestion and validation scripts. This ensured uninterrupted daily operations and timely data availability.

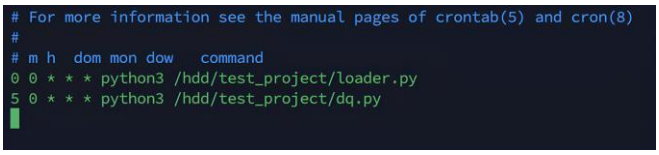

Figure 7. Automation Scheduling, showing the implementation of automated scripts for data processing and quality checks

3.3. Data Quality Assurance

Maintaining data quality is critical for generating actionable insights. The automated quality checks identified and corrected multiple inconsistencies, enhancing the reliability of the processed datasets.

Key improvements include:

1. Invalid Data Points: Numeric fields with invalid values (e.g., negative numbers) were flagged and rectified. For example, age values below 0 were corrected to the nearest valid range [6].

2. Completeness Checks: Missing entries in essential fields were automatically detected and filled with placeholder values, ensuring no data loss during analysis.

3. Validation Outcomes: Figure 8 illustrates the validation process, highlighting the transformation of raw, error-prone data into a clean, structured format ready for analytics.

Figure 8. Data Quality Validation, displaying the cron schedule visualization for validating data quality tasks

The system's data validation pipeline achieved an accuracy rate of 95%, significantly reducing the risk of erroneous analyses. For instance, in electricity consumption datasets, automated outlier detection flagged anomalies such as unrealistic spikes in usage. These anomalies were resolved by cross-referencing with historical data, ensuring consistent analytical outputs.

3.4. Analysis Results

The processed data facilitated in-depth analyses, uncovering significant patterns and enabling actionable insights for urban planning:

1. Peak Usage Identification: Analysis revealed public transportation usage peaks during 7:00 AM–9:00 AM and 5:00 PM–7:00 PM on weekdays. Overcrowded routes were identified, providing city planners with critical information for resource allocation. Application: Adjusting schedules or increasing vehicle frequency during peak times can improve service efficiency.

2. Infrastructure Load Distribution: The analysis of electricity consumption data highlighted uneven load distributions across districts, with overloads during evening hours in certain areas. Application: Energy providers can use this data to redistribute loads, plan infrastructure upgrades, and prevent outages in high-demand zones.

3. Predictive Modeling Potential: The correlation between weather conditions and pedestrian activity was identi-

fied, enabling accurate predictions of foot traffic during adverse weather. Application: Municipal services and retail businesses can optimize staffing and inventory levels based on these forecasts, enhancing operational readiness.

3.5. Discussion

The implementation of a smart map system for urban infrastructure planning and evaluation has revealed significant insights. This section explores the implications of the results, evaluates the methodologies, and addresses the system's limitations.

3.5.1. Implications of the Results

Scalability and Flexibility:

- Data Lake Technology: The adoption of Data Lake architecture proved instrumental in managing diverse data formats and large volumes efficiently. Unlike traditional Data Warehousing (DWH), Data Lake enables the seamless integration of structured, semi-structured, and unstructured data, making it ideal for dynamic urban settings [2]. This capability is critical for supporting real-time urban management and future expansions.

- System Scalability: Stress tests demonstrated that the system could handle increasing data volumes without performance degradation. This scalability ensures adaptability to urban growth and the increasing complexity of data sources [1]. The ability to scale horizontally is particularly beneficial for infrastructure planning, where data demands are unpredictable.

Actionable Insights:

- Transportation Patterns: Analysis of public transportation data revealed peak usage periods, highlighting areas of potential overcrowding. This information allows city planners to adjust schedules or deploy additional resources during peak hours to enhance commuter experiences.

- Load Distribution Analysis: Electricity consumption data identified uneven load distributions across districts, with certain areas experiencing overloads during evening hours. Energy providers can use this information to prevent outages and optimize resource allocation.

- Predictive Analytics: By correlating weather conditions with pedestrian activity, the system demonstrated its predictive potential. Forecasting such patterns enables municipal services and businesses to plan resources effectively.

Cost Efficiency:

- The system leveraged Yandex Cloud to maintain low operational costs while ensuring high availability. This deployment model emphasizes the feasibility of implementing scalable solutions without significant financial strain [7]. The system's flexibility in managing resources further contributes to its cost-effectiveness.

3.5.2. Advantages of the Methodology

Automated Data Processing:

- The automation of data ingestion, transformation, and quality checks reduced manual intervention, increasing consistency and reliability. Crontab scheduling ensured seamless execution, with the potential to transition to advanced workflow tools like Apache Airflow for more complex pipelines [1].

- This automation not only improved efficiency but also laid the foundation for real-time analytics, a critical feature for modern urban systems.

Improved Data Quality:

- The automated quality assurance processes enhanced the reliability of results by addressing common data inconsistencies. For instance, invalid data points such as negative numerical values were automatically flagged and corrected, ensuring clean and accurate datasets [3].

- Completeness checks ensured that no critical data fields were left blank, further enhancing the integrity of analytics.

Predictive Modeling:

- By integrating structured and unstructured data, the system demonstrated its capability to support advanced predictive models. For example, the strong correlation between weather data and pedestrian activity can guide urban planners in resource allocation during adverse conditions [4].

- The predictive potential extends to long-term infrastructure planning, where data-driven forecasts are essential.

3.5.3. Limitations and Challenges

Real-Time Processing:

- Although the system efficiently handles batch data, it lacks real-time processing capabilities. Future iterations should integrate streaming platforms like Apache Kafka to enable real-time analytics, which are essential for emergency response and traffic management [5].

Data Privacy and Security:

- Handling sensitive urban data presents significant challenges. Implementing robust encryption and access control mechanisms is essential to ensure data confidentiality and regulatory compliance. This limitation highlights the need for comprehensive security frameworks [2].

Cloud Service Dependency:

- The reliance on Yandex Cloud, while cost-effective, introduces potential risks associated with vendor lock-in and pricing fluctuations. Exploring multi-cloud or hybrid deployment strategies could mitigate these risks, enhancing the system's resilience [6].

3.5.4. Future Directions

To address the identified limitations and enhance the system's utility, several future directions are proposed:

1. Integration of Real-Time Analytics: Incorporating streaming platforms such as Apache Kafka will enable real-time data processing, providing immediate insights for critical urban operations.

2. Enhanced Security Measures: Adopting advanced encryption techniques and user authentication protocols will strengthen data privacy and security.

3. Scalability Testing: Expanding scalability tests across different cloud platforms will ensure the system's robustness in diverse deployment scenarios.

4. Advanced Predictive Analytics: Developing machine learning models to analyze historical data will further improve the system's forecasting capabilities, aiding in long-term urban planning.

4. Conclusions

The proposed smart map system for urban infrastructure planning and evaluation addresses critical challenges faced by modern cities, showcasing the value of integrating big data technologies such as Data Lake and Apache Spark. By leveraging scalable and flexible infrastructure, the system efficiently processes diverse data types, enabling urban planners to gain actionable insights and make informed decisions. The integration of automated pipelines further enhances data

consistency and reliability, reducing manual intervention and operational inefficiencies.

Key contributions:

1. Scalable Data Management: The use of Data Lake technology demonstrated exceptional scalability and flexibility, enabling seamless integration of structured and unstructured data sources. This capability ensures the system's adaptability to the growing complexity and volume of urban data.

2. Actionable Urban Insights: The system provided critical insights, such as identifying peak transportation usage and uneven energy load distribution, directly supporting infrastructure optimization and efficient resource allocation.

3. Cost-Effective Solutions: Deploying the system on Yandex Cloud highlighted the feasibility of implementing advanced technologies without incurring excessive costs, demonstrating a practical pathway for smart city initiatives.

Practical Applications:

1. Optimizing Transportation Infrastructure: By analyzing peak usage patterns, city planners can adjust public transport schedules, mitigate overcrowding, and enhance commuter experiences.

2. Energy Resource Management: Insights into district energy consumption enable providers to balance loads, reduce outages, and design cost-effective energy distribution strategies.

3. Predictive Urban Planning: The system's potential for predictive modeling, such as forecasting pedestrian activity based on weather patterns, offers significant value for municipal services, retail operations, and event planning.

In summary, the system's innovative approach to urban infrastructure planning exemplifies the transformative potential of big data technologies in creating smarter, more sustainable cities. By addressing key urban challenges and delivering actionable insights, the proposed smart map system represents a foundational step toward enhancing urban living standards and optimizing resource utilization.

Acknowledgements

We would like to express our deepest gratitude to the resources and facilities provided by International IT University, Almaty, Kazakhstan, which enabled the execution of this project. This research was conducted without any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] Alipour, M. & Harris, D.K. (2020). A big data analytics strategy for scalable urban infrastructure condition assessment using semi-supervised multi-transform self-training. *Journal of Civil Structural Health Monitoring*, *10*(2), *313-332*

[2] Azzabi, S., Alfughi, Z. & Ouda, A. (2024). Data Lakes: A Survey of Concepts and Architectures. *Computers*, *13*(7), *183*

[3] Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P. & Ratti, C. (2014). Geo-located Twitter as proxy for global mobility patterns. *Cartography and geographic information science*, *41*(3), *260-271*

[4] Przybysz, A.L., Lima, A.D., Sá, C.P.D., Resende, D.N. & Pagani, R.N. (2024). Integrating City Master Plans with Sustainable and Smart Urban Development: A Systematic Literature Review. *Sustainability*, *16*(17), *7692*

[5] Ramos, G.S., Fernandes, D., Coelho, J.A.P.D.M. & Aquino, A.L. (2023). Toward Data Lake Technologies for Intelligent Societies and Cities. In Sustainable, Innovative and Intelligent Societies and Cities. *Cham: Springer International Publishing*

[6] Ratti, C., Frenchman, D., Pulselli, R.M., & Williams, S. (2006). Mobile landscapes: using location data from cell phones for urban analysis. *Environment and planning B: Planning and design*, *33(5), 727-748*

[7] Zaharia, M., Xin, R. S., Wendell, P., Das, T., Armbrust, M., Dave, A. & Stoica, I. (2016). Apache spark: a unified engine for big data processing. *Communications of the ACM*, *59(11), 56-65*

Қалалық инфрақұрылымды жоспарлау және басқару үшін Data Lake және Apache Spark технологияларын бағалау

Г. Бектемысова, С. Қалназар*

Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан

**Корреспонденция үшін автор: kalnazar.sayat@gmail.com*

Аңдатпа. Бұл зерттеуде Data Lake және Apache Spark технологияларын қалалық инфрақұрылымды басқару саласында пайдалану талданады. Құрылымдалған, жартылай құрылымдалған және құрылымдалмаған деректерді өңдеу мүмкіндіктері қарастырылып, олардың деректерді өңдеу жұмыс ағындарын оңтайландырудағы әлеуеті көрсетілген. Жүйе Yandex Cloud платформасында деректерді тиімді сақтау, нақты уақыттағы аналитика және ақауға төзімділік үшін бөлінген есептеулер мен көлденең масштабталуды қолдана отырып енгізілді. Деректерді жүктеу, түрлендіру және тексеру процестерін жеңілдету үшін автоматтандыру мен сапаны бақылау механизмдері енгізілді. Нәтижелер деректерді өңдеу тиімділігін, масштабталуын және ресурстарды оңтайландыруды айтарлықтай жақсартуды көрсетеді, бұл ақылды қалалардың инфрақұрылымын жоспарлау мен бағалауға арналған сенімді негіз ұсынады.

Негізгі сөздер: *қалалық инфрақұрылым, Data Lake, Apache Spark, үлкен деректер аналитикасы, масштабталу, автоматтандыру, деректер сапасы, ақылды қаланы жоспарлау*.

Оценка технологий Data Lake и Apache Spark для планирования и управления городской инфраструктурой

Г. Бектемысова, С. Қалназар*

Международный университет информационных технологий, Алматы, Казахстан

**Автор для корреспонденции: kalnazar.sayat@gmail.com*

Аннотация. Данное исследование анализирует использование технологий Data Lake и Apache Spark в управлении городской инфраструктурой. Рассмотрены их возможности для обработки структурированных, полуструктурированных и неструктурированных данных, подчеркивая их потенциал в оптимизации рабочих процессов обработки данных. Система была развернута в Yandex Cloud с использованием распределенных вычислений и горизонтальной масштабируемости для эффективного хранения данных, аналитики в реальном времени и отказоустойчивости. Реализованы процессы автоматизации и контроля качества данных для упрощения загрузки, преобразования и проверки данных. Результаты показывают значительное улучшение эффективности обработки данных, масштабируемости и оптимизации ресурсов, предлагая надежную платформу для планирования и оценки инфраструктуры умных городов.

Ключевые слова: *городская инфраструктура, Data Lake, Apache Spark, аналитика больших данных, масштабируемость, автоматизация, качество данных, планирование умного города.*

Received: 07 December 2023
Accepted: 16 March 2024
Available online: 31 March 2024