Integrating Data Lakes with Self-Adaptive Serious Games

Michalis Pingos[a], Spyros Loizou[b] and Andreas S. Andreou[c]
Department of Computer Engineering and Informatics, Cyprus University of Technology, Limassol, Cyprus

Keywords: Big Data, Data Lakes, Serious Games, Self-Adaptation, Learning Optimization, Data Processing, Big Data Analytics.

Abstract: Big Data challenges traditional data management tools due to their size, speed of generation, and variety of formats. The application of Big Data has become essential in areas like serious games, where it enhances functionality and effectiveness. Serious games benefit significantly from Big Data analytics, allowing for real-time data collection, processing and analysis of a vast number of users/players and their interactions. Traditional data management systems strive to handle the complexity of Big Data, particularly in environments like serious games, where diverse data sources create heterogeneity. This paper presents an approach that employs Data Lakes and semantic annotation as a solution for providing a scalable, flexible storage system for raw and unstructured data, enabling real-time processing and efficient management of Big Data produced by serious games. The effectiveness of the proposed approach is demonstrated through a speech therapy game example developed for the purposes of this study. A qualitative evaluation and comparison with two rival approaches is also performed using a set of criteria introduced in this work. The proposed approach offers an effective solution for handling data in multi-user gaming environments thus enhancing adaptability, personalization, and functional flexibility of serious games, and driving better user engagement and outcomes.

1 INTRODUCTION

Big Data refers to extremely large and complex datasets that cannot be easily managed, processed, or analysed using traditional data management tools and methods (Yaqoob et al., 2016). It is not just about the size of the data, it also involves the speed at which it is generated and the variety of formats it comes in, which brings heterogeneity and complexity. Given this complexity and scale, the application of Big Data has become increasingly crucial in fields such as serious games, where its integration enhances both functionality and effectiveness (Nasrollahi et al., 2023).

Serious games, designed for education, training, healthcare, or simulations, increasingly rely on Big Data to ensure they are effective and meaningful (Pérez et al., 2023). Unlike traditional games focused on entertainment, serious games aim to teach, train, or solve real-world problems. The integration of Big Data analytics in serious games allow for real-time collection, processing, and analysis of vast and complex datasets. This enhances the ability to personalize experiences, track user progress, and evaluate outcomes comprehensively (Zhao et al., 2023). By leveraging Big Data, developers and researchers can validate the effectiveness of serious games, optimize learning experiences, and ensure the games achieve their intended impact on real-world problems (Kosmides et al., 2018).

As Big Data continues to evolve in terms of volume, velocity, and variety, traditional data management systems struggle to handle its scale and complexity (Al-Mansour, 2023). This is especially evident in fields like serious games, where diverse and rapidly generated datasets from player interactions, behavioural analytics, and multimedia inputs create a highly heterogeneous environment (Hooshyar et al., 2018). Data Lakes have emerged as a solution to address these challenges by providing a centralized repository that can store raw,

[a] https://orcid.org/0000-0001-6293-6478
[b] https://orcid.org/0009-0009-3433-3245
[c] https://orcid.org/0000-0001-7104-2097

Pingos, M., Loizou, S. and Andreou, A. S.
Integrating Data Lakes with Self-Adaptive Serious Games.
DOI: 10.5220/0013467700003928
In *Proceedings of the 20th International Conference on Evaluation of Novel Approaches to Software Engineering (ENASE 2025)*, pages 762-772
ISBN: 978-989-758-742-9; ISSN: 2184-4895

unstructured, and semi-structured data in its native format. Unlike traditional databases that require predefined schemas, Data Lakes offer the flexibility and scalability necessary to accommodate the multidimensional nature of Big Data, thereby enabling real-time processing and deeper insights (Hai et al., 2023) critical to the development and enhancement of serious games.

This paper proposes a novel approach that utilizes Data Lakes as an effective approach to managing Big Data in serious games enhancing user experience. The effectiveness of this approach is demonstrated using a speech therapy game example presenting how it may improve serious games functionality, adaptability, and personalization by extracting actionable insights that lead to more engaging and effective gaming environments.

The rest of the paper is structured as follows: Section 2 outlines the technical background and related work in the area of Big Data, serious games and Data Lakes. Section 3 presents the proposed approach and associated methodology, along with a speech therapy game scenario example. Preliminary validation conducted to assess the effectiveness of this approach is showcased in Section 4. Experimentation conducted to assess performance is provided in Section 5. Finally, Section 6 summarizes the paper's findings and proposes directions for future research exploration.

2 LITERATURE AND TECHNICAL OVERVIEW

As mentioned in the previous section, integrating heterogeneous Big Data sources while maintaining consistency and accuracy is a complicated task. The increased volume, velocity, and complexity of data introduce performance and scalability issues, necessitating robust computational resources and efficient storage architectures.

Data Lakes, a term coined by Pentaho's CTO James Dixon in 2010 (Nolte & Wieder, 2022), offer a scalable solution to the challenge of storing vast amounts of raw, unprocessed data in its native format. By providing a centralized repository that can accommodate data in various formats from disparate and heterogeneous sources, Data Lakes enable efficient storage, processing, and real-time analysis. This ensures that performance and scalability demands are met without compromising data integrity or accessibility (Gorelik, 2019).

Fang et al. (2015) discuss the growing trend of Data Lakes as a method for organizing and utilizing large datasets. They explore the concept of Data Lakes, their benefits in addressing Big Data challenges, and their relationship with existing enterprise data warehouses. The authors also address common concerns about Data Lakes and offer practical guidance on developing a Data Lake strategy and integrating it into a company's data management architecture. Furthermore, their paper leverages low-cost technologies like Apache Hadoop to improve the capture, refinement, archival, and exploration of raw data within an enterprise.

Olawoyin et al. (2021) present an innovative Data Lake architecture tailored for managing and processing large-scale Arctic datasets. The approach introduced in that paper emphasizes open access, scalability, and compatibility with machine learning tools, enabling researchers to analyse diverse data types like satellite imagery and climate records.

Data Lakes need to be supported by a strong metadata mechanism to offer a scalable solution, which must involve tagging and mapping of the content of data sources. This ensures that data remains organized and meaningful, preventing the lake from turning into a swamp (Cherradi & El Haddadi, 2021). Specifically, the metadata mechanism should provide sources' information, data characteristics, and access and governance details.

Sawadogo and Darmont (2021) provide a comprehensive overview of Data Lake concepts, architectures, and metadata management techniques. Their work addresses the challenges of managing large volumes of heterogeneous data, focusing on architectural considerations and the crucial role of metadata in preventing data swamps. The authors also review existing Data Lake definitions, explore various architectures and technologies for implementation, and analyze metadata management strategies.

As an example, Beheshti et al. (2018) introduce a system for managing knowledge graphs within a Data Lake environment. CoreKG offers services for knowledge graph storage, retrieval, and manipulation, enabling efficient querying and analysis of interconnected data. The system aims to bridge the gap between Data Lakes and knowledge graphs, facilitating knowledge discovery and informed decision-making by leveraging the rich semantic information contained within knowledge graphs.

The authors in Pingos and Andreou (2022) propose a novel framework for enhancing data discovery and understanding within Data Lakes by

leveraging metadata semantics. The authors introduce the concept of blueprints, which capture semantic relationships between metadata elements, enabling more effective querying and analysis of data. By exploiting this semantic enrichment of the Data Lake, that framework aims to improve data management and facilitate knowledge extraction from complex datasets within Data Lake environments. The concept of the Semantic Data Blueprints introduced in that work is also utilized in the present paper to enrich the Data Lake part for supporting serious games.

The global serious games market is projected to grow significantly, with an estimated market size of USD 14.06 billion in 2024, expected to reach USD 43.65 billion by 2029. This represents a compound annual growth rate (CAGR) of 25.43% during the forecast period from 2024 to 2029 (Mordor Intelligence, 2024). According to SkyQuest Technology (2021), serious games revolutionize learning and training by combining gameplay with real-world applications across various field such as healthcare, defence and education, making training engaging and practical. Big Data and advanced technologies, such as virtual reality (VR), augmented reality (AR), and artificial intelligence (AI) are enhancing their impact and reshaping the future of skill development.

While the combination of serious games and Big Data offers exciting possibilities, several challenges need to be addressed. Serious games, especially those with complex simulations or multiplayer interactions, can generate massive datasets from various sources, such as player inputs, sensor data, and external databases (Pérez et al., 2023). Traditional data storage and processing methods struggle to handle the volume and velocity of this data (Marz & Warren, 2015). Efficient and scalable solutions are needed to manage this massive amount of information.

According to Caggianese et al. (2018), Big Data analysis requires significant computational resources, which can be costly and complex to manage. To address this issue, the authors present a tele-rehabilitation system leveraging serious games and cloud-based data analytics. Designed for post-stroke patients, the system collects real-time motor data and provides a decision support service for analysis. This data, along with reports and recommendations, is made available on the cloud, enabling clinicians to remotely assess rehabilitation progress and personalize therapies. The study also includes a pilot study and qualitative analysis of the system's clinical impact and acceptance.

The exploration of how learning analytics can enhance serious games is investigated in Alonso-Fernández et al., (2018). The authors analyse Big Data gathered from a serious game designed to teach computer programming, focusing on player behaviour and learning outcomes. By examining the relationship between in-game actions and learning gains, they identify key insights for improving game design and personalization. The study provides valuable insights for developers and data scientists, enabling the creation of more effective games through the personalization of gamification in serious games.

Taking into account the current literature and the challenges that have emerged in this field, a new approach utilising Data Lakes is proposed in this paper as an effective solution for managing Big Data to enhance serious game experience. The concept of the approach and its applicability are demonstrated through a speech therapy example showcasing how this Big Data architecture can enhance the serious gaming experiences and provide meaningful insights.

3 PROPOSED APPROACH

As outlined earlier, the proposed approach integrates Data Lakes enriched with a Semantic Data Blueprint (SDB) to address the challenges of Big Data in serious gaming environments, enabling enhanced functionality, adaptability, and personalization. The approach is designed to process diverse data inputs from multiple sources (users, environments) in serious games and transform them into actionable insights that improve user experience.

Diverse gaming data sources are fed into the Data Lake. These sources include structured data (e.g., SQL tables), semi-structured data (e.g., JSON or XML), and unstructured data (e.g., text, voice recordings, or images). The diversity of these inputs reflects the heterogeneous data produced from serious gaming environments. The data management architecture of the proposed approach is a Data Lake, which serves as a scalable and flexible storage system for managing raw, unstructured, semi-structured, and structured data as presented in Figure 1.

The data generated within the serious game, including user interactions, progress metrics, feedback, real-time assessments etc., is pushed directly into the Data Lake in its raw format. This ensures that diverse and heterogeneous data is ingested without requiring prior transformation or structuring, allowing for real-time integration and future processing. The corresponding data is stored in the Data Lake which is structured using the pond and puddle architecture introduced in Pingos and Andreou (2022). Each pond includes data for each user, while each puddle refers and points to the

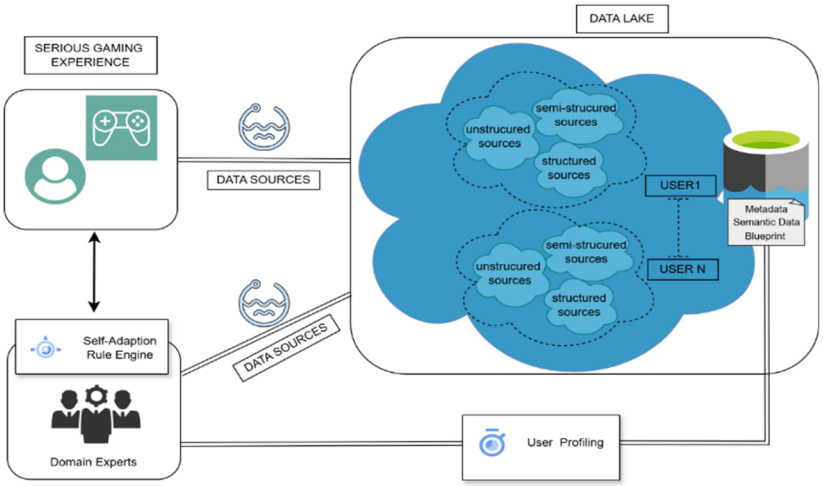

Figure 1. Data Lake Integration for Adaptive and Personalized Serious Gaming Experience.

unstructured, semi-structured and structured data produced from a particular user during the game. On the other hand, a dedicated SDB complements the Data Lake (right part of Figure 1), which enriches the stored data by adding layers of metadata and semantic organization. The SDB ensures that the data within the Data Lake is not only stored efficiently, but it is also contextualized and made accessible for advanced analysis. This semantic mechanism allows the system to perform complex queries, identify patterns, and extract meaningful insights from raw data, facilitating the creation of personalized gaming experiences. By leveraging the SDB, this approach provides the ability to create detailed user profiles that capture preferences, progress, and challenges. These profiles could feed back into the serious gaming environment, allowing for real-time adjustments, personalized content delivery, and enhanced engagement.

The demonstration example utilized in this work comes from the speech therapy domain (see Section 5), and generates a wide variety of data through user interaction with therapy tools, such as speech recognition applications or therapist-guided activities. These inputs include: (i) unstructured data, like audio recordings of pronunciation exercises, where users practice specific phonemes or words; (ii) semi-structured data, such as therapist notes and feedback captured in formats like JSON or XML; and, (iii) structured data, including assessment scores and progress metrics that quantify the user's improvement over time. This diversity reflects the complexity of speech therapy data, providing a rich foundation for deeper analysis and personalized therapy enhancements (Riad et al., 2022).

```
<UserInteractionData>

  <StructuredData>
          <Example>
              <UserID>U123</UserID>
              <AccuracyScore>92</AccuracyScore>
              <GameLevel>3</GameLevel>
               <Date>2024-12-05T14:30:00Z</Date>
          </Example>
          <SemanticTagging>
              <Tag>/progress/advanced</Tag>
              <Tag>/phoneme/medial</Tag>
          </SemanticTagging>
   </StructuredData>

  <SemiStructuredData>
          <Example>
              <PhonemeSelection>
                  <Phoneme>p</Phoneme>
                  <Phoneme>b</Phoneme>
              </PhonemeSelection>
              <DifLevel>Intermediate</DifLevel>
          </Example>
          <SemanticTagging>
              <Tag>/phoneme</Tag>
              <Tag>/difficulty/intermediate</Tag>
          </SemanticTagging>
    </SemiStructuredData>

  <UnstructuredData>
          <Example>
              <AudioFile>U123lvl3.wav</AudioFile>
          </Example>
          <Metadata>
              <AudioFormat>WAV</AudioFormat>
              <Duration>3.2s</Duration>
              <Phoneme>p</Phoneme>
          </Metadata>
          <SemanticTagging>
              <Tag>/phoneme/initial</Tag>
          </SemanticTagging>
      </UnstructuredData>

</UserInteractionData>
```

Figure 2. A Semantic Data Blueprint example in XML.

A serious game called “Phoneme Quest” was developed in this work as a demo example. This educational activity was designed to help practice phoneme isolation by identifying initial, medial, and final sounds in words. In this game, players match picture cards based on shared phonemes, fostering phonemic awareness and supporting language development (Vihman, 2017). According to the proposed approach, each user has their own dedicated Data Lake Pond, designed to store and organize their

data. Within each pond, data is further divided into smaller parts called puddles for more granular organization: One puddle organizes structured data like accuracy scores, level completions, and performance metrics; another puddle is used for semi-structured data, such as therapist-configured settings (e.g., phoneme selection, difficulty levels stored in formats like JSON or XML); and, a third puddle address unstructured data, including audio recordings of speech attempts and visual feedback from gameplay. This pond-and-puddle approach provides a clear, scalable approach for organizing and analysing user data at multiple levels of granularity, enabling tailored insights and targeted therapy improvements for each individual leveraging Big Data (Pingos & Andreou, 2022).

The SDB plays a vital role for the Data Lake while introducing metadata, semantic tagging, and data linking to transform raw data into actionable insights for serious gaming environments. An example of an SDB in XML format is provided in Figure 2 describing user interaction data of User123. Each interaction added to the SDB of the Data Lake provides semantic tagging with attribute values for the structured, unstructured and semi-structured data generated by the user in every interaction.

Note that every time the specific or other user interacts with the game, the data produced is pushed as is to the corresponding data pond and puddle, while the associated information is added to the metadata of the SDB as presented earlier in Figure 1. Another reason for dividing user data into distinct puddles based on the data type is to streamline and standardize processing methods when needed. This approach ensures that similar types of data can be processed uniformly, increasing efficiency and reducing complexity, while the SDB provides the ability to handle efficiently heterogeneous data sources, a defining characteristic of Big Data.

The self-adaptation rule engine is a core component of the proposed approach (see Figure 1) that facilitates the dynamic adjustment of gameplay to meet user needs. The engine leverages input provided by domain experts, data collected from users during game sessions, and analytical models to perform adaptation of game complexity, difficulty, and scenarios based on user performance and learning goals. Domain experts play a vital role in designing and guiding the adaptation process through a set of rules that essentially describe when and where adaptation should be performed in real-time. Their input ensures that the game's structure, goals, and scenarios align with the educational needs of the users. Based on these rules, the engine can temporarily adjust specific game elements, such as reducing or increasing the number of challenges or simplifying tasks, to align with player capabilities. For example, experts may define specific visual or auditory elements to be displayed on screen suitable for players with unique conditions (e.g., Down syndrome), such as calming images for stress reduction or tailored challenges according to the learning goals (e.g. thickening lines or removing colours). Additionally, domain experts may set rules describing that when a user fails to identify the correct answer from a list of options presented, then the engine modifies gameplay by offering hints. The rules are documented using a standardized approach, such as XML or JSON files, and the self-adaption engine parses these files to make decisions.

The SDB supports both the self-adaptation rule engine and the domain experts by organizing and structuring game-related data. This metadata-driven approach ensures that player interactions, performance metrics, and game scenarios are stored efficiently in the Data Lake. Domain experts utilize this semantically rich data to validate and redefine game goals, while the rule engine uses it to adapt gameplay in real time. This collaboration between the SDB, domain experts, and the rule engine ensures that the game remains aligned with its intended objectives, while providing an engaging and adaptive environment for the user.

The architecture of the Data Lake introduced here allows for real-time ingestion of data from numerous users simultaneously, ensuring that user interactions, gameplay metrics, and other inputs are stored as is without latency. For example, when multiple players participate in the "Phoneme Quest" serious game, their performance data is immediately sent to the structured puddles of their respective ponds and is stored in the Data Lake in its native format, including the enrichment of the metadata semantic blueprint. The same approach applies, for example, to audio recordings and therapists' feedback, which are stored

```
PREFIX game: <http://example.org/game#>
PREFIX meta: <http://example.org/meta#>

SELECT ?userID ?phoneme ?successRate ?timeSpent
WHERE {
    ?interaction
        game:userID       ?userID ;
        game:phoneme      ?phoneme ;
        game:successRate  ?successRate ;
        game:timeSpent    ?timeSpent .

    FILTER (?successRate >= 90 && ?timeSpent <= 10)
}
```

Figure 3. SPARQL query to extract user performance with a success rate of 90 percent or higher and time spent of 10 minutes or less

in unstructured and semi-structured formats respectively, along with updates to the SDB. This core concept of the Data Lake helps serious games systems to handle large amounts of diverse data efficiently, providing the means for self-adaptation through efficient retrieval (querying) and processing of this data that supports dynamic adjustments in gameplay, real-time analytics, and personalized experiences.

Querying the SDB is more effective and efficient than querying raw data because it provides enriched context, semantic relationships, and integration of heterogeneous data types. The SDB actually organizes raw, structured, semi-structured, and unstructured data with metadata and semantic tagging, enabling the construction and execution of complex queries that directly link user interactions, gameplay metrics, and session information. For example, querying the SDB using SPARQL allows quick and direct extraction of actionable insights, such as identifying users who struggled with a specific phoneme while correlating this difficulty with session timestamps and therapists' feedback. In contrast, querying raw data requires significant preprocessing, including parsing the raw data, something which increases complexity and delays insights.

The SDB provides an essential layer of semantic metadata, which makes the raw data actionable. Specifically, the SDB can link user data points together, such as a phoneme practiced, the corresponding success rate, and the time spent on the exercise. This metadata enriches the data stored in the puddles, allowing queries that combine heterogeneous data types. For example, as presented in Figure 3, a SPARQL query can be designed to extract information about users who achieved a 90% success rate on a specific phoneme, while spending less than 10 minutes on the corresponding exercises.

The integration of Data Lakes and the SDB within the proposed approach provides a robust solution for managing the complexity and scale of data generated in serious gaming environments. By organizing user data into ponds and puddles, the approach achieves granular storage and streamlined processing, ensuring scalability and efficiency. Simultaneously, the SDB enhances the data's utility by adding semantic layers and metadata, transforming raw data into actionable insights. This combination not only addresses the challenge of handling heterogeneous data, but also facilitates advanced analytics and personalized experiences. The effectiveness of this approach is further demonstrated in the upcoming experimentation sections, which details its practical application and evaluates its impact on real-world use cases.

4 PRELIMINARY EVALUATION

First, a qualitative evaluation of the proposed approach for managing Big Data in serious gaming environments is conducted. To this end, it is essential to compare it with existing works that address similar challenges. Two notable studies in this domain are Learning Mechanics − Game Mechanics (LM-GM), and SGA.

LM-GM is a conceptual model designed to integrate learning objectives seamlessly into the gameplay of serious games by aligning educational activities with game features. This alignment ensures that gameplay actively promotes educational outcomes, fostering engagement while achieving learning goals. For example, problem-solving, as a learning mechanic, can be mapped to puzzle-solving as a game mechanic to teach critical thinking skills (Arnab et al., 2015).

On the other hand, SGA is a data-driven approach designed to collect, analyse, and interpret the data generated by serious games to evaluate and enhance their effectiveness. It focuses on leveraging player interaction data, such as gameplay behaviours, decision-making patterns, and performance metrics, to assess learning outcomes, engagement, and skill development. By utilizing techniques from learning analytics and educational data mining, SGA provides insights into player progress, identifies areas of difficulty, and informs the design of personalized learning experiences. While highly effective for post-hoc analysis, SGA is less focused on real-time processing or integrating diverse data types, making it best suited for structured analysis of gameplay data (Shute et al., 2013).

The present paper defined the following five criteria to facilitate a qualitative evaluation of LM-GM and SGA and a comparison with the proposed approach: *Data Handling Capabilities*, *Semantic Enrichment*, *Personalization*, *Scalability*, and *Integration Flexibility*. Furthermore, ratings such as *Low*, *Medium*, and *High* for each criterion were set to streamline the qualitive comparison and to simplify the evaluation process.

Data Handling Capabilities refers to the ability to manage and process the diverse types of data generated in serious games, such as structured data, semi-structured data and unstructured. A *Low* rating indicates lack of the ability to handle such diverse data types, wh*ium* value suggests that the model can manage some data types, typically structured data, but may face challenges with semi-structured or unstructured data, often relying on additional tools to bridge these gaps. A *High* rating, on the other hand,

signifies a comprehensive ability to efficiently process all types of data with mechanisms for seamless and scalable management, making it suitable for modern, data-intensive serious gaming environments.

Semantic Enrichment assesses the capacity to enhance raw data by adding context and meaning through semantic layers, metadata tagging, and organized contextual frameworks. A *Low* rating indicates lack of any semantic enrichment features or contextual organization abilities, making it difficult to derive advanced insights from raw data. A *Medium* value suggests that the model provides basic metadata tagging or limited semantic organization, but it does not include advanced functionality like semantic querying or data linking. *High* signifies robust semantic abilities, including support for ontologies and/or knowledge graphs, advanced metadata systems, and the ability to perform complex semantic queries, enabling deeper insights and enhanced contextual understanding of the data.

Personalization Support evaluates the capacity to deliver tailored user experiences based on interaction data, preferences and progress information. A *Low* rating indicates that the approach lacks mechanisms for personalization and does not analyse user-specific data to customize the gaming experience, while a *Medium* rating means that it offers limited support, such as enabling post-hoc personalization where analytics are used to design tailored content after gameplay. A *High* value reflects advanced capabilities, where the model is able to leverage data insights so as to dynamically adjust game elements through the self-adaption rule engine, like content, difficulty levels, or pacing to match the user's needs, creating a more engaging and individualized experience.

Scalability assesses the capacity to manage increasing demands as data volume, user numbers, or game complexity grow, without compromising performance. A *Low* rating indicates that the approach is not equipped to handle large-scale data or multiple users effectively, resulting in performance bottlenecks under high demand. A *Medium* value suggests moderate *Scalability*, where the model can accommodate some growth, but may face challenges with very large datasets, complex data structures, or simultaneous user interactions. A *High* rating, however, represents robust *Scalability*, where the approach seamlessly manages massive datasets, diverse data types, and large user bases across varying gaming environments, maintaining optimal performance.

Finally, *Integration Flexibility* assesses the ability to work seamlessly with various external tools, platforms, and data sources, which is critical in complex and diverse serious game environments. A *Low* rating indicates a rigid model with limited or no capability to integrate with external systems, restricting adaptability and scalability. A *Medium* value signifies moderate integration capabilities, where the model can connect with some tools or data sources, but may require extensive customization or additional effort. A *High* rating, on the other hand, reflects a highly flexible approach designed for seamless interoperability, allowing it to accommodate diverse technologies and environments efficiently.

The aforementioned evaluation criteria are considered fundamental for approaches handling Big Data, as they collectively ensure the organization, analysis, and application of diverse datasets. These elements enable models to adapt to growing demands, integrate seamlessly with various tools, and deliver meaningful insights. Together, they can form the backbone of effective, scalable, and adaptable systems in heterogeneous data-intensive environments. Therefore, this set of criteria serve as a strong qualitative evaluation basis for the comparison of the three approaches, the proposed one and two similar or close rivals.

The approach introduced in this paper achieves a *High* rating in *Data Handling Capabilities* due to its ability to manage structured, semi-structured, and unstructured data efficiently using a Data Lake architecture and the semantic data blueprint. In contrast, LM-GM is rated *Low* as it focuses on conceptual mapping of learning and game mechanics, lacking functionality for handling or processing data. SGA receives a *Medium* rating because it effectively manages structured data, such as player scores or performance metrics, but struggles with semi-structured or unstructured data, which are increasingly prevalent in modern serious gaming environments.

LM-GM may be rated with a Low value for Semantic Enrichment because it does not incorporate features for adding context or meaning to raw data, focusing on aligning learning and game mechanics conceptually instead. Similarly, SGA is also rated Low, as it emphasizes data analytics, but lacks capabilities for semantic enrichment, such as advanced metadata tagging or ontology support. In contrast, the proposed approach is rated High due to its robust SDB, which integrates advanced metadata tagging, ontology support, and semantic querying capabilities. This enables the model to add context and meaning to raw data in the Data Lake, facilitating deeper analysis and actionable insights.

Table 1. Evaluation and comparison of the approaches.

Criteria	Proposed Approach	LM-GM	SGA
Data Handling Capabilities	High	Low	Medium
Semantic Enrichment	High	Low	Low
Personalization Support	High	Low	Medium
Scalability	High	Low	Medium
Integration Flexibility	High	Low	Medium

In addition, the proposed approach could be characterized with *High Personalization*. Creating user profiles in the Data Lake enables the dynamic adaptation of the game content, difficulty levels, or other elements to enhance individual engagement and learning outcomes. LM-GM is rated *Low* as it lacks mechanisms for *Personalization* and focuses mainly on game design principles. SGA is rated *Medium* because it provides preliminary insights into player performance, which can be used to inform personalization strategies.

On the other hand, LM-GM may be rated with *Low Scalability* as it is not designed to handle large datasets, multiple users, or complex data environments. Similarly, SGA is rated with *Medium Scalability* as it performs sufficiently with structured data but faces challenges when dealing with very large datasets, users, or highly complex data scenarios. In contrast, the proposed approach achieves a *High* rating due to its robust design and Data Lake support, which efficiently handle large-scale, heterogeneous data, making it highly suitable for the growing data demands of serious games.

Finally, the approach presented in this work may be characterized with *High Integration Flexibility*, as it supports seamless interoperability with various data sources, tools, and platforms (for example the Hadoop ecosystem). This ensures that is may easily adapt to different technological ecosystems and gaming environments. On the other hand, LM-GM may be rated with *Low* for this criterion since it is purely a theoretical framework and does not support integration with external systems. SGA can be rated as *Medium* due to the fact that although it integrates well with specific tools, it has limited flexibility and requires customization to work with diverse data sources or platforms.

Table 1 summarizes the comparison between the three serious gaming approaches discussed in this section. The analysis highlights that the proposed approach significantly outperforms the two others based on the selected criteria.

5 EXPERIMENTAL EVALUATION

5.1 Design of Experiments

The experiments conducted serve two primary objectives. First, they aimed to evaluate the ability of the proposed method to integrate Data Lakes with Self-Adaptive Serious Games, leveraging Semantic Data Blueprints. Second, the experiments focused on assessing the performance and effectiveness of the approach in terms of data volume. This section outlines the rationale behind the design and execution of a set of experiments to meet the above objectives. Although this experimentation is limited it can offer some preliminary useful insights.

The "Phoneme Quest" serious game was utilized for generating the experimental data. Initially, a Data Lake metadata mechanism was constructed for this game, which was also uploaded on GitHub (https://github.com/mfpingos/DL-Self-Adaptive-Serious-Games). The DL metadata semantic data blueprint is described using an XML file, which captures user interactions stored in the Data Lake as presented earlier in Figure 1. Initially, a basis data set was constructed having one user interacting with the game. Next, Python scripts were utilized to enlarge the number of users/interactions thus creating a synthetic dataset able to assess performance and efficiency as the volume of data increases.

Every time a user interacts with the game a set of metadata is produced in XML format as shown earlier in Figure 2. This metadata includes structured, semi-structured and structured information, something very common in serious games interaction. In the example of Figure 2 the data captures a detailed interaction between a user and the game focusing on phonetic learning or a speech-related task. More specifically, the structured data provides numerical information and metrics, such as a unique User ID ("U123"), an Accuracy Score of 92, and a Game Level of 3, all linked to the date and time of interaction. This format allows for clear tracking of user progress and performance in the system, which is essential for monitoring user engagement or measuring success in a task-oriented environment. The semantic tags, such as "/progress/advanced" and "/phoneme/medial", categorize the user's activities, offering insights into their focus areas, like phonemes

related to vowels and their progress level within the system.

The semi-structured data offers more flexible information, such as the specific phonemes the user engaged with (“/p/” and “/b/”), as well as a Difficulty Level of “Advanced”. This type of data allows for a clearer understanding of the user's interaction, indicating that the user is working on advanced-level tasks involving phonetic sounds.

The unstructured data includes an audio file (“U123lvl.wav”), which contains speech content tied to the user's learning process. The metadata attached to this file includes the duration of 3.2 seconds and the audio format (WAV), along with the tagging of “/phoneme/advanced”, highlighting the content's focus on advanced phonetic tasks. Together, these data types provide a comprehensive view of user interactions, from performance scores and task difficulty to the specific phonetic elements being exercised and the media used to engage the user.

The experimentation consists of two parts, each executing a SPARQL query:

(i) The execution of a query on the Data Lake metadata resulting in information about users’ interactions with game level (difficulty) equal or greater than 2 (medium):

```
PREFIX ns: <http://example.org/ns#>
SELECT ?userID ?gameLevel
WHERE {
    ?interaction ns:userID ?userID .
    ?interaction ns:gameLevel
        ?gameLevel .
    FILTER(?gameLevel >= 2)
       }
```

(ii) The execution of a query gathering information for a specific user and its accuracy score for each interaction:

```
PREFIX ns: <http://example.org/ns#>
SELECT ?accuracy Score ?date
WHERE {
    ? interaction ns:userID
        "specific_user" .
    ? interaction ns:accuracyScore
        ?accuracyScore .
    ?interaction ns:date ?date .
       }
```

The increase in the number of users and interactions directly impacts the size of the respective XML file, a crucial element parsed to extract users that match a query. In the first experiment the XML file describing 1 user resulted in a size of 0.002MB, 10 users of 0.01MB and 100 users of 0.078MB. Each of the files contains one interaction of each user. For the second experimentation, the metadata describing two interactions with the serious game of 1 user resulted in a size of 0.002MB, 10 users of 0.02MB and 100 users of 0.216MB.

The experimental setup utilized a server configured with three Virtual Machines (VMs). Each VM was provisioned with 4 CPU cores, while the host server itself contained a total of 48 cores. Memory allocation per VM was 8192MB, and each had an 80GB hard disk capacity. The software stack comprised Hadoop 3.3.6 for distributed computing, Python 2.7.5 for scripting tasks, and synthetic datasets generated from the “Phoneme Quest” serious game. Apache Jena was integrated into the system for handling SPARQL queries.

5.2 Experimental Results

The two benchmark SPARQL queries were executed 100 times for each synthetic metadata to facilitate a comprehensive comparative analysis. As mentioned in the previous section and presented in Tables 2 and 3, the metadata size and the number of users differ for each run.

Table 2 shows the time required for querying the metadata mechanism using SPARQL(i) requesting information among all users and the number of results returned (i.e., the number of user interactions). This table also summarizes the average execution time, metadata size, and the number of results returned when executing SPARQL(i) queries across varying numbers of users. For a single user with a metadata size of 0.002 MB, the average execution time is 0.0005 seconds, returning results for 1 user. With 10 users and a metadata size of 0.01 MB, the average execution time increases to 0.001 seconds, returning results for 10 users. For 100 users, the metadata size rises significantly to 0.078 MB, with an average execution time of 0.02 seconds, returning results for 90 users. As can be observed, when the number of users increases, both the metadata size and execution time show also a noticeable increase. This increase is more profound at the beginning (10 users) for the average execution time, but it then stabilizes at a lower level. Additionally, the results in Table 2 illustrate the computational overhead caused by increasing users and metadata size. While the execution time grows moderately with additional users, the increase in metadata size is more substantial, reflecting the resource intensiveness of managing larger datasets. The results underline the scalability challenges in handling larger user bases and metadata volumes while maintaining acceptable query performance.

Table 2. Average execution time (after 100 executions) requesting information for different number of users with different sizes of metadata running the SPARQL(i) query.

Number of users	Metadata Size	Average running time	Number of user interactions returned
1 User	0,002 MB	0,0005s	1 user
10 Users	0,01 MB	0,0010s	10 users
100 Users	0,078 MB	0,0020s	90 users

Table 3. Average execution time (after 100 executions) requesting information for one user among many users' interactions with different sizes of metadata running the SPARQL(ii) query.

Number of users and interactions	Metadata Size	Average execution time
1 User (2 interactions)	0.002 MB	0,0011s
10 Users (20 interactions)	0.02 MB	0,0075s
100 Users (200 interactions)	0.216 MB	0,0100s

The second experiment is presented in Table 3, and shows the time required for querying the metadata mechanism using SPARQL(ii) query, which requests information for one particular user. Table 3 summarizes the average execution time, metadata size, and the number of results returned when executing SPARQL(ii) queries across varying numbers of users and serious game interactions when requesting information for a specific user. For a single user with 2 interactions and metadata size of 0.002 MB, the average execution time is 0.0011 seconds. For 10 users with 200 interactions the metadata reaches a size of 0.01 MB and the average execution time increases to 0.0075 seconds. Finally, for 100 users with 200 interactions, the metadata size rises proportionally to 0.078 MB, with an average execution time of 0.01 seconds. Furthermore, Table 3 presents the computational overhead caused by increasing users and interactions within the metadata mechanism. While the execution time shows a gradual increase with the addition of new users and their interactions, the growth in metadata size is significantly more pronounced.

The tables in each benchmark illustrates the computational efficiency of querying metadata using the SPARQL mechanism compared to querying the actual data. While execution time grows moderately with an increase in users, and interactions and metadata size increase substantially, the overhead introduced remains manageable. A direct comparison with querying the actual data would not be feasible unless a dedicated query is constructed and executed depending of the type of the data (RDBMS, images etc). Therefore, such a comparison would not be directly performed, at least on the same comparison level. This demonstrates that querying metadata is significantly more efficient and scalable than directly querying the actual data, which would involve much larger datasets and higher number of queries, and, therefore, would result in substantially longer execution times. The results emphasize the advantage of the proposed approach and the ability to provide self-adaptive serious games experience with multiple users and large volumes of data via the integration of Data Lakes.

6 CONCLUSIONS & FUTURE WORK

This paper presented a new approach for managing Big Data in serious game environments based on Data Lakes and Semantic Data Blueprints. The target was to offer a data storage architecture able to host data produced at high speeds and volumes in its raw format and then retrieve it effectively to process it so as to provide improvements in game experience, customization, and adaptability.

The approach offers efficient game data management and real-time processing, which address the challenge of heterogeneous data sources in serious games. Integration of various data types, such as structured, semi-structured and unstructured, into Data Lakes and description or tagging of this data with semantic layers provided the means for utilizing historical data to yield deeper performance insights and personalized gaming experience through the self-adaption rule engine. The approach introduced was demonstrated using an example of a speech therapy game and was evaluated and compared against two close rivals using a set of five quality features that suggested its superiority. Experimentation was also conducted through two benchmark datasets to assess performance and effectiveness of the proposed approach as data volume increased with very satisfactory results.

Future work will focus on the following: (i) Expand the application of the proposed approach in more application areas targeting at supporting education and training and performing further testing and evaluation with real-world experimentation. (ii) Investigate the extent to which real-time data processing and adaptation impact system latency and

computational resource requirements. (iii) Explore how optimization strategies for parallel execution in a multi-VM environment could enhance scalability and resource utilization and provide further experimentation. (iv) Employ advanced analytics and recommender systems to provide more standardized user profiling and extend the self-adaptive capabilities of serious games by dynamically adjusting content, scenarios and gameplay based on user behaviour and experiences.

REFERENCES

Yaqoob, I., Hashem, I. A. T., Gani, A., Mokhtar, S., Ahmed, E., Anuar, N. B., & Vasilakos, A. V. (2016). Big data: From beginning to future. International Journal of Information Management, 36(6), 1231-1247.

Nasrollahi, H., Lampropoulos, I., Werning, S., Belinskiy, A., Fijnheer, J. D., Veltkamp, R. C., & van Sark, W. (2023). Review of Serious Energy Games: Objectives, Approaches, Applications, Data Integration, and Performance Assessment. Energies, 16(19), 6948.

Pérez, J., Castro, M., & López, G. (2023). Serious games and ai: Challenges and opportunities for computational social science. IEEE Access, 11, 62051-62061. assessment, and improvement, 101-134.

Zhao, Y., Gao, W., & Ku, S. (2023). Optimization of the game improvement and data analysis model for the early childhood education major via deep learning. *Scientific reports*, *13*(1), 20273.

Kosmides, P., Demestichas, K., Adamopoulou, E., Koutsouris, N., Oikonomidis, Y., & De Luca, V. (2018, August). Inlife: Combining real life with serious games using iot. In 2018 IEEE Conference on Computational Intelligence and Games (CIG) (pp. 1-7). IEEE.

Al-Mansour, F. (2023). IoT Data Management: Challenges and Solutions in Handling Vast Amounts of Information. *Innovative Computer Sciences Journal*, *9*(1).

Hooshyar, D., Yousefi, M., & Lim, H. (2018). Data-driven approaches to game player modelling: a systematic literature review. ACM Computing Surveys (CSUR), 50(6), 1-19.

Nolte, H., & Wieder, P. (2022). Realising data-centric scientific workflows with provenance-capturing on data lakes. Data Intelligence, 4(2), 426-438.

Hai, R., Koutras, C., Quix, C., & Jarke, M. (2023). Data lakes: A survey of functions and systems. *IEEE Transactions on Knowledge and Data Engineering*, *35*(12), 12571-12590.

Gorelik, A. (2019). The enterprise big data lake: Delivering the promise of big data and data science. O'Reilly Media.

Fang, H. (2015, June). Managing data lakes in big data era: What's a data lake and why has it became popular in data management ecosystem. In *2015 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER)* (pp. 820-824). IEEE.

Olawoyin, A. M., Leung, C. K., & Cuzzocrea, A. (2021, December). Open data lake to support machine learning on Arctic big data. In 2021 IEEE International Conference on Big Data (Big Data) (pp. 5215-5224). IEEE.

Cherradi, M., & EL Haddadi, A. (2021, November). Data lakes: A survey paper. In *The Proceedings of the International Conference on Smart City Applications* (pp. 823-835). Cham: Springer International Publishing.

Sawadogo, P., & Darmont, J. (2021). On data lake architectures and metadata management. *Journal of Intelligent Information Systems*, *56*(1), 97-120.

Beheshti, Amin, Boualem Benatallah, Reza Nouri, and Alireza Tabebordbar. "CoreKG: a knowledge lake service." *Proceedings of the VLDB Endowment* 11, no. 12 (2018): 1942-1945.

Pingos, M., & Andreou, A. S. (2022, April). Exploiting Metadata Semantics in Data Lakes Using Blueprints. In *International Conference on Evaluation of Novel Approaches to Software Engineering* (pp. 220-242). Cham: Springer Nature Switzerland.

SkyQuest Technology. (2021). Global serious games market size, share, growth analysis by gaming platform and application – Industry forecast 2023-2030. SkyQuest. https://www.skyquestt.com/report/serious-games-market

Warren, J., & Marz, N. (2015). *Big Data: Principles and best practices of scalable realtime data systems*. Simon and Schuster.

Caggianese, G., Cuomo, S., Esposito, M., Franceschini, M., Gallo, L., Infarinato, F., ... & Romano, P. (2018). Serious games and in-cloud data analytics for the virtualization and personalization of rehabilitation treatments. IEEE Transactions on Industrial Informatics, 15(1), 517-526.

Alonso-Fernández, C., Perez-Colado, I., Freire, M., Martínez-Ortiz, I., & Fernández-Manjón, B. (2018). Improving serious games analyzing learning analytics data: Lessons learned. Games and Learning Alliance Conference, 287-296. Cham: Springer.

Riad, R., Ali, Y., Tits, N., Humbert, S., & Dutoit, T. (2022). KSoF: A Dataset of Fluency and Stuttering Annotations of German Speech. arXiv. https://arxiv.org/abs/2203.05383

Vihman, M. M. (2017). Learning words and learning sounds: Advances in language development. British Journal of Psychology, 108(1), 1-27.

Arnab, S., Lim, T., Carvalho, M. B., Bellotti, F., Freitas, S. de, Louchart, S., … De Gloria, A. (2015). Mapping learning and game mechanics for serious games analysis. British Journal of Educational Technology, 46(2), 391–411. https://doi.org/10.1111/bjet.12113

Shute, V. J., Ventura, M., Bauer, M., & Zapata-Rivera, D. (2013). Measuring and Supporting Learning in Games: Stealth Assessment. The MIT Press.