SemiStructQuest: Unveiling NoSQL Database Concepts Through Gamified Learning Platform Analyses*

Nelly Barret[1,*,†], Mélina Verger[1,†]

[1]INSA Lyon, CNRS, Universite Claude Bernard Lyon 1, LIRIS, UMR5205, 69621 Villeurbanne, France

Abstract

This research proposal aims to develop an educational tool that supports the learning of the complex and diverse underlying concepts in the NoSQL database domain. The proposed gamified learning platform is meant to complement traditional instruction and will enable us to identify effective learning pathways from students' interactions. Exercises and game levels will be designed with a focus group of domain experts and higher-education instructors. Next, interaction traces and self-reported background information will both be collected and further analyzed to determine how learners' profiles relate to the sequences of exercises they follow. Domain experts will then assess the pedagogical relevance of these sequences to derive insights that can inform teaching practices and support future adaptive guidance for learners. Next steps include organizing the initial focus group, developing the platform prototype, and evaluating its usability and usefulness through further user-centered evaluations.

Keywords

data systems education, NoSQL databases, sequence mining

1. Introduction

Data management, and in particular NoSQL databases, are at the core of various educational programmes, including computer science, but also law, finance, or business, and are necessary for many careers. Therefore, it is highly important to accompany students in the learning of this complex topic, which interconnects various concepts, including data, models, and queries.

Despite most STEM courses include courses on NoSQL databases, there has been little research on educational tools to facilitate the student learning. Indeed, traditional learning (lectures, written exercises, and hand-on projects) is still predominant in these courses. However, this approach is often inefficient because NoSQL databases may take various forms (document-based, graph, column-based, etc) and are inherently complex to understand (and to teach). These observations push toward the need of designing novel gamified tools to engage students in the learning of this complex topic and to help them decipher the intricacies of the various NoSQL databases. Additionally, understanding how students progress and where they struggle is an important factor both for teachers and learners. In turn, this allows providing insights to educators on how to apprehend the teaching of these topics. Students also benefit from these insights because they can revise misunderstood concepts and can better identify where to put efforts. Finally, taking into account learner profiles helps drawing conclusions on effective methods and personalizing the learning within tools.

In this research proposal paper, our main objective is to **design a gamified learning platform to identify effective learning pathways of NoSQL database concepts from users' interactions, in correlation with learner background information**. This work thus lies at the intersection of three research fields: data management (DM), learning analytics (LAK) and human-computer interaction (HCI). The proposed platform, named SemiStructQuest, will complement traditional teaching approaches

DataEd'26: 5th International Workshop on Data Systems Education

*Published in the Proceedings of the Workshops of the EDBT/ICDT 2026 Joint Conference (March 24-27, 2026), Tampere, Finland

*Corresponding author.

†These authors contributed equally.

nelly.barret@insa-lyon.fr (N. Barret); melina.verger@insa-lyon.fr (M. Verger)

0000-0002-3469-4149 (N. Barret); 0000-0002-5839-882X (M. Verger)

(including courses and labs) while providing insights to learners and teachers. To reach this objective, our research proposal is a 3-step process:

1. First, we aim to **define a set of exercises** for the gamified platform with the collaboration of a focus group. The set of exercises seeks to be diverse, covering the set of notions seen during lectures, and of various difficulties of NoSQL database concepts. The goal is that students can play with various exercises depending on their skills, time, and engagement. This task happens upstream, before the creation of the platform, as shown by the top-right blue rectangles in Figure 1. This step is detailed in Section 3.
2. The second step is to **collect traces** from students interacting with the platform. This happens as long as the students play with the exercises of the platform, as shown by the first orange rectangle below the box in Figure 1. Moreover, learners may fill background questionnaires before starting with the platform (green rectangles at the bottom left of Figure 1). These data will be used jointly for deeper analysis. The trace and profile collection is explained in Section 4.
3. Finally, **learning sequences are mined** from the collected traces, as shown by the last orange rectangles at the bottom right of Figure 1. We further **align these sequences with learner background profiles**. This two-stage task allows to extract insights on the learning of NoSQL databases, as detailed in Section 5. We also open discussion regarding this task in Section 6, before concluding in Section 7.

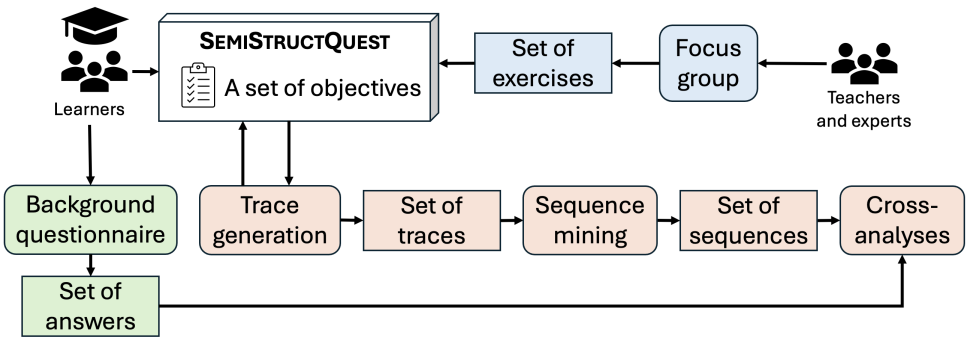


Figure 1: The workflow to extract and analyze learning sequences from SemiStructQuest traces. Rounded rectangles are tasks and straight ones are task outputs. Lines draw their connections.

2. Related work

In this section, we discuss existing tools for teaching NoSQL databases as well as strategies to gamify pedagogical tools. Finally, we position our paper with regards to this existing literature.

NoSQL educational tools. Several surveys [1] and educational tools, such as TriQL [2], NoSQL e-learning laboratory [3], or PrairieLearn [4], have been proposed recently in this domain. TriQL [2] helps students to learn how to query SQL, MongoDB and Neo4J databases by providing an interface to build queries (selecting tables, attributes, conditions, etc.). Then, TriQL transposes users' queries into an intermediate Datalog query before generating three queries, respectively in SQL, MongoDB and Cypher. The NoSQL e-learning laboratory [3] makes the query writing more interactive by connecting to Moodle, a widely-used learning platform, with MongoDB and CouchDB databases. Then, students can issue queries against theses databases, but also view the data model of the databases. Finally, PrairieLearn [4] focuses on providing various exercise sets to students until they reach mastery and on providing instant feedback about their answers (in terms of positive and negative points). A recent survey [1] emphasizes the lack of thorough surveys dealing with NoSQL database education despite an increasing growth of NoSQL usage both in academia and industry. Moreover, the survey highlights

the diversity of concepts learned during NoSQL courses, including data models and modeling, queries, scalability and optimization, as well as security. This variety of topics makes the teaching of NoSQL databases even more complex.

Gamification. Educational tools may take various forms, including intelligent tutoring systems (ITS) and game-based learning (GBL) platforms [5]. Both aim to propose to learners series of exercises that follow given pedagogical objectives. These systems also often provide feedback to and/or help learners based on their answers. Thus, they complement traditional teaching methods, such as lectures, written exercises, and textbooks. Also, a key focus of educational tools is learner engagement. They often offer varied, gamified exercises featuring rewards and objectives, leading to an efficient "learning-by-doing" approach. Furthermore, one of the crucial aspects of these tools is their ability to collect the traces produced by learners as they interact with the system. Such interaction traces are very helpful for better understanding how learners apprehend exercises and how they process information, offering insights into the learning process between pedagogical notions.

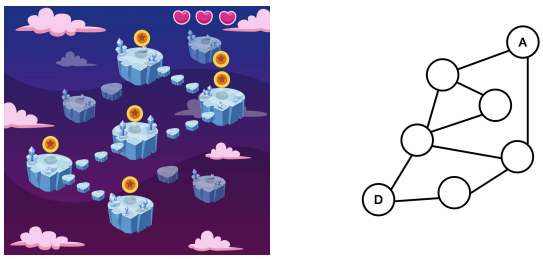
Positioning. While our contribution builds on the insights and lessons learnt from existing literature, our proposal differs on the following points. First, our gamified platform will rely on pathways individualized to each learner based on their ability to progress and learn NoSQL databases (see Sections 3). This improves upon existing similar tools, such as the NoSQL e-learning lab [3] and Duolinguo, which impose a fixed pathway for every learner (the learner has to retry until success). Second, we analyze these individual pathways to extract insights from the student interactions with our tool. To the best of our knowledge, analyzing student traces from a sequence mining point of view is novel given the related work mentioned earlier. Third and last, while profiling students is a well-known technique to gather complementary information or provide personalized pathways [6], we take a different approach by combining self-reported and log data to find correlations between students' profiles and their pathways.

3. Building a gamified learning platform

In this section, we present our approach to build SemiStructQuest, the gamified learning platform, composed of game levels (Section 3.1) and gamified elements (Section 3.2). Illustrative examples will be given in the following.

3.1. Design of the game levels

The game levels will be designed in close collaboration with domain experts. Indeed, historically there are three key types of models that are commonly integrated into educational tools: the domain model (which outlines the relationships between concepts), the learner model (which tracks the learner's progress), and the pedagogy model (which governs the functionality of the educational tool) [7]. In our



(a) (b)

Figure 2: Illustrations of (a) an example of a game map (source: Freepik) and (b) an example of level paths.

case, while our learner model will be based on analyzing declarative and log data, and our pedagogy model will be based on gamified elements (further explained in the next section), we will build the domain model specifically for NoSQL database education by relying on expert knowledge.

To do so, we will organize a focus group, which is a widespread research method from the field of HCI [8], composed of up to ten domain experts and higher education instructors specialized in NoSQL databases. The objective is to (i) collect qualitative feedback to develop the game levels [8] and to (ii) gather meta-knowledge on each game level proposed, such as the level of difficulty, the database concepts involved, and quality criteria regarding the expected answers (as different answers may be correct for a given question).

More precisely, we will ask the focus group to identify a set of representative exercises, similar to those that might appear in a NoSQL database course. Next, participants will be asked to link each exercise to the relevant NoSQL database concept(s), the associated pedagogical objective(s), and levels of difficulty. For example, an exercise might involve the concept of filtering a collection in MongoDB with the objective of writing valid MongoDB queries at a beginner level. Finally, they will come up with some quality criteria for the possible correct queries, taking into account their flexibility and diversity. Therefore, this process will allow us to design corresponding game levels and the scoring approach, detailed in the next section.

3.2. The gamified learning platform

The goal of this part is to describe the gamified elements considered to engage learners in NoSQL database education within our platform. In addition to these gamified elements, the platform includes a non-gamified element: the questionnaire, that can be filled at launch of the platform, preceding the access to the overall game map, described below.

Game map. The idea behind the game map is to translate and materialize the outcomes from the focus group into game levels, as shown in Figure 2a. Additionally and importantly, unlike platforms such as Duolingo [9], learners can choose multiple paths to complete the game, as in Figure 2b. Indeed, the aim is not to impose a linear sequence of activities but to let learners select the NoSQL database notions they wish to engage with, in a playful setting. This decision builds on two motivations: one for gamification and one for analysis. First, the gamified design is intended to encourage users to explore the different levels of the platform and play at their own pace. Second, by not enforcing a predetermined sequence, we can analyze, from multiple users' interactions, what would be the most pedagogically relevant learning pathways, i.e., the game level sequences most frequently played. These emerging sequences would then be reviewed by the domain experts to obtain feedback on their relevance and thus provide insights that could inform traditional teaching practices, as further explained in Sections 5 and 6.

Interactive interface. Once users select a game level from the game map, they will enter an interactive interface designed for writing queries, inspired by the NoSQL e-learning lab [3]. Each level is accompanied by a problem to solve, and the platform will be connected to a database server, allowing users to query data directly. Users can make as many queries as needed to complete the game level (the scoring system is explained later on). If they choose to leave a level before completing it, they will lose a life heart (see below), but they can follow another level path to explore in hopes of reaching the final game level.

Gamified elements. In addition to the game map, several other gamified elements will be incorporated. First, each game level's difficulty will be represented by one to three stars, which also reflect the potential reward for completing that level (see Figure 2a). Scoring details are provided below.

Second, players will have three lives, represented by heart symbols, to complete the entire game. This system is designed to motivate users to engage with levels that match their self-assessed mastery level, encouraging them to challenge themselves without feeling overwhelmed.

Scoring. The scoring system will be cumulative across the game levels. Each level will have a maximum score, determined in collaboration with the focus group, and the level's difficulty (indicated by stars) will serve as a multiplication factor. To earn a score, two conditions must be met: first, the player must complete the level by finding a secret word hidden in the database; second, the quality of the queries will be evaluated based on predefined criteria established with the focus group. The more criteria a query satisfies, the higher the score awarded.

4. Collecting traces and profiles

In this section, we show how to collect interaction traces (Section 4.1) and learner profiles (Section 4.2) from the SemiStructQuest platform, both valuable data for better understanding how students interact with the platform and what could be the most relevant pedagogical sequences for learning NoSQL database concepts (Section 5).

4.1. Platform traces

The platform will timely store the traces produced by the learners while interacting with the interface and exercises. When a learner starts a new session, the platform will record: his/her anonymized identifier, each exercise that the learner starts, the duration to succeed with the number of tries for each of them, and the obtained score (recall the previous section on scoring). Such information allows to build **learning sequences** for each learner, i.e., the ordered set of exercises that the student decides to do (recall that SemiStructQuest is a pathway-free tool where learners may decide at any time which exercise they want to work on). Learning sequences as defined in the paper contain the minimal yet sufficient information to perform further learning sequence mining (see Section 5) toward reaching the goal of aligning learning sequences and learner background profiles.

Note that many other traces could be collected from the platform, such as the number of clicks in the interface, eye-tracking, the usage of external resources, etc. While they would allow to have finer-grained analyses of learner's behaviors and progress, they are much more complex to exploit in conjunction of sequence mining and may be considered for future work.

4.2. Learner profiles

To complement the traces collected from the platform and provide deeper analyses, questionnaires are a relevant method to collect learner background profiles. Additionally to SemiStructQuest, we defined the following questions, which will be treated as binary and ordinal variables in the analyses (Section 5):

- "Have you ever taken a course on NoSQL databases in your education (either mandatory or optional)?" with the answers being "Yes" or "No".
- "Have you ever already used any kind of NoSQL databases?" with the answers being "Yes" or "No".
- "Have you ever already written queries for any kind of NoSQL databases?" with the answers being "Yes" or "No".
- "What is your self-assessed proficiency level with NoSQL databases?" with the answers being a 5-levels Likert scale [10], i.e., "Novice" (1), "Advanced beginner" (2), "Competent" (3), "Proficient" (4), or "Expert" (5).

The questionnaire will be accessible when entering the platform for the first time. In turn, its completion will be optional, allowing users to decide whether to provide background information. Note that, as for the traces collection (Section 4.1), this questionnaire is intentionally minimal yet sufficient for our purpose. Indeed, when aligned with the interaction traces, it will allow us to determine whether learners with different backgrounds employ different learning strategies, as explored in Section 5.

5. Analyzing traces and learner profiles

In this section, we detail how to extract insights from the traces and questionnaires collected from SemiStructQuest. First, we focus on determining a set of frequent learning strategies from the traces (Section 5.1). Next, we show how to align these learning strategies with the learner profiles (Section 5.2).

5.1. Finding common and uncommon learning sequences

In order to make sense of all the learning sequences collected by the platform, we apply *sequential pattern mining* algorithms to **extract the most and least frequent learning (sub-)sequences**. Most frequent ones represent ordered sets of exercises that multiple learners do more often than others, and conversely.

Sequential pattern mining algorithms take as input a set of sequences and outputs the most, or least, frequent sequences or sub-sequences in the input set. In turn, they may find that some sub-sequences are very frequent (rather than a complete sequence involving all the proposed exercises) because these algorithms work with all the sub-sequences of the input sequences. For instance, with 4 exercises denoted E_1, E_2, E_3 and E_4, learning sequences produced by the learners could be $E_1 \rightarrow E_2 \rightarrow E_3 \rightarrow E_4$, or $E_1 \rightarrow E_2 \rightarrow E_4 \rightarrow E_3$, or any other combination (the sequence may also be a more complex path, as shown in Figure 2b, if the learner explores various exercises before completing them). In this example, there is no most frequent learning sequence, but one can observe that the sub-sequence $E_1 \rightarrow E_2$ appears in both sequences. This generalizes to the observation that, as long as the number of exercises and/or learners increase, the harder it is to find most frequent sequences but the easier it is to find most frequent sub-sequences. Similarly, we extract least frequent patterns in the obtained learning sequences. By doing so, we may reveal uncommon learning strategies, an important marker for the lack of understanding or different experiences than others with SemiStructQuest.

In practice, there exist two main families of sequential pattern mining algorithms: *Apriori*-based techniques [11] or *Pattern-Growth* [12]. In a nutshell, both families of algorithms are able to find the same set of most, or least, frequent patterns (sequences or sub-sequences) for a given set of sequences but rely on different underlying data representations. In any case, these algorithms require to set the *minimum support* S, a real value in $]0, 1]$. A high S requires the pattern to be very frequent in the input sequences (high popularity), and conversely. When implementing our platform, we will adopt Apriori for its ease of implementation and expect to set $S = 0.5$, i.e., the pattern should appear in at least 50% of the collected sequences.

5.2. Aligning learning sequences with learner profiles

From the frequent learning sequences extracted as previously explained, we seek to relate these sequences with the background information reported from the users. By applying pattern mining algorithms on the sequences, we seek to identify **"optimal learning strategies"**, those that appear most effective with respect to both game-level success and learning progress. More precisely, these strategies correspond to sequences of exercises that are associated with higher success rates in the game as well as improvements in learners' performance over time. For instance, novice or beginner learners (recall the background questionnaire in Section 4.2) are expected to focus on easy and fast exercises first, before starting with more complex ones. In contrast, advanced or expert learners are expected to start with intermediate or difficult exercise first, as they already know and understood the concepts proposed in initial exercises.

To determine and visualize these correlations, we will compute a **confusion matrix**, or a heatmap, where the x-axis represents the number of occurrences of the identified sequences (whether as the most frequent or infrequent), and the y-axis represents the questionnaire variables. All questionnaire variables are binary except for the Likert-scale item, which will be converted to a one-hot encoded set of variables. We will then compute the Spearman correlation coefficient between each questionnaire variable and the number of occurrences of each identified sequence. The resulting heatmap will allow us

to identify clusters of correlations that can help interpret the sequences in light of learners' background information and provide insights into relevant teaching practices and learning pathways.

6. Discussion

Before concluding this proposal paper, we discuss two aspects of our paper, namely the computation of common pathways and the implementation of the proposed platform.

Common pathways. Our proposal seeks to identify frequent learning pathways within the platform logs. Primarily, extracting frequent pathways will bring valuable feedback on how students engage with the tool and the designed game levels. However, frequent patterns may also reflect common struggles or misconceptions because students often go in the wrong direction before coming back. To distinguish between frequent and effective pathways in our joint analyses (recall Section 5.2), we will take into account quantitative measures of the students success with the scores they earn for completing levels with respect to the level difficulty.

Implementation and test. This paper describes a proposal toward providing a novel tool for the education of NoSQL databases. Implementing it in a prototype will be the first step, before providing a more comprehensive software. Implementation will be made open-source and easily deployable in various environments to ease usage and tests in real NoSQL database courses.

7. Conclusions and perspectives

In this paper, we present our research proposal to build a gamified learning platform for NoSQL database education. This platform aims to provide a gamified learning experience that complements traditional teaching in this challenging domain, while also allowing us to identify effective learning pathways from users' interactions. The exercises and game levels will be designed in collaboration with a focus group composed of domain experts and higher education instructors specializing in NoSQL databases.

Then, we aim to collect both interaction traces and self-reported data from a background questionnaire to perform cross-analyses. From the interaction traces, we intend to identify sequences of exercises that may reveal useful pedagogical insights. The self-reported data will help us determine whether certain sequences, and thus certain strategies, are favored depending on learners' background, prior knowledge, and familiarity with NoSQL databases.

In order to extract insights from the analyzed sequences, we will cross-validate them with the learners' background information, before letting domain experts reviewing them to assess their pedagogical relevance. This could even inform traditional teaching practices on how to proceed with the teaching of NoSQL databases, notably in terms of pedagogical sequences. Additionally, these insights could be used in the future to provide adaptive guidance to learners through feedback on the platform. For example, a learner with little or no prior background might receive feedback such as: "You did not successfully complete this exercise; you can try this alternative exercise".

Our next steps are therefore to organize the focus group, develop the platform, and prepare the modules for collecting interaction traces and self-reported data. Once the prototype is ready, a key concern will be the platform interface, particularly with respect to two dimensions: usability and usefulness. These HCI aspects will likely be examined with a second focus group, including non-experts such as users.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] N. Tripathi, Nosql database education: A review of models, tools and teaching methods, Journal of Systems and Software (2025) 112391.

[2] A. Alawini, P. Rao, L. Zhou, L. Kang, P.-C. Ho, Teaching data models with triql, 2022, pp. 16–21. doi:10.1145/3531072.3535320.

[3] A. Werner, M. Bach, Nosql e-learning laboratory—interactive querying of mongodb and couchdb and their conversion to a relational database, in: International Conference on Man–Machine Interactions, Springer, 2017, pp. 581–592.

[4] M. West, G. L. Herman, C. Zilles, Prairielearn: Mastery-based online problem solving with adaptive scoring and recommendations driven by machine learning, in: 2015 ASEE Annual Conference & Exposition, 2015, pp. 26–1238.

[5] A. I. Abdul Jabbar, P. Felicia, Gameplay engagement and learning in game-based learning: A systematic review, Review of educational research 85 (2015) 740–779.

[6] M. Lefevre, S. Jean-Daubias, N. Guin, Personnaliser des séquences de travail à partir de profils d'apprenants, in: EIAH 2009-Environnements Informatiques pour l'Apprentissage Humain, 2009, p. 4.

[7] B. P. Woolf, Building intelligent interactive tutors: Student-centered strategies for revolutionizing e-learning, Morgan Kaufmann, 2010.

[8] Research Methods for Human-Computer Interaction, Cambridge University Press, 2008.

[9] M. Shortt, S. Tilak, I. Kuznetcova, B. Martens, B. Akinkuolie, Gamification in mobile-assisted language learning: a systematic review of duolingo literature from public release of 2012 to early 2020, Computer Assisted Language Learning 36 (2023) 517–554. URL: https://doi.org/10.1080/09588221.2021.1933540.

[10] R. Likert, A technique for the measurement of attitudes., Archives of psychology (1932).

[11] R. Agrawal, T. Imieliński, A. Swami, Mining association rules between sets of items in large databases, in: Proceedings of the 1993 ACM SIGMOD international conference on Management of data, 1993, pp. 207–216.

[12] J. Han, J. Pei, Y. Yin, Mining frequent patterns without candidate generation, ACM sigmod record 29 (2000) 1–12.