

VR-ETL: Leveraging Virtual Reality for Pre-Transformation ETL Data Validation

Maja Cerjan¹, Kornelije Rabuzin¹, and Leo Mršić²

¹ Faculty of Organization and Informatics, University of Zagreb,
Pavlinska 2, Varaždin, Croatia,
Department of Theoretical and Applied Foundations of Information Sciences
macerjan@foi.unizg.hr (corresponding author),
krabuzin@foi.unizg.hr

² Algebra Bernays University,
Gradišćanska 24, 10000 Zagreb, Croatia
leo.mrsic@algebra.hr

Abstract. Data validation in ETL workflows is most often performed after transformation, even though many data quality issues originate earlier in the pipeline. In practice, this means that errors are detected late, when tracing their source and applying corrections already requires additional effort. In this work, we explore whether virtual reality (VR) could be used as an alternative interface for addressing this problem earlier, at the stage of raw data inspection. We present VR-ETL, a proof-of-concept system that allows users to inspect and correct tabular data before transformation by interacting with it inside an immersive 3D environment. The system imports CSV datasets and applies a set of rule-based validation checks targeting common data quality problems, such as missing values, format inconsistencies, and basic referential integrity violations. Detected issues are visualized directly within the 3D representation of the data, using color-coded cells and simple spatial summaries that reflect error distribution across columns and records. A functional prototype was implemented using Unity and OpenXR and tested on structured datasets containing up to 10,000 records. During controlled technical testing, the system supported automated validation, real-time visual feedback, and immediate re-validation after user corrections, while maintaining performance suitable for immersive use on consumer-grade hardware. All intentionally seeded rule-based errors were detected. Although usability and workflow efficiency were not evaluated in this study, the prototype demonstrates that pre-transformation data validation in VR is technically feasible and provides a concrete starting point for future empirical investigations.

Keywords: Virtual Reality, ETL, Error Handling, Data Validation, Immersive Analytics.

1. Introduction

In recent years, the importance of data-driven decision-making for finance, healthcare, and business intelligence operations of many organizations has received a lot of attention [16]. The worth of these findings depends on how well analysts process data and how

well the first set of records was recorded. The process of data quality maintenance has progressed from basic relational database restrictions into an intricate problem which handles multiple data formats and origins [5]. The evaluation of unprocessed data becomes unreliable because these datasets contain missing information together with formatting problems and structural irregularities [28]. Organizations with extensive operations implement Extract, Transform, and Load (ETL) frameworks to manage their data integration and cleaning operations [38,30]. The frameworks have achieved technological readiness but data quality management stands as the main barrier which prevents their implementation. The ETL process follows a standard error detection approach, which happens after transformation steps and because of this it requires longer procedures to track and fix records which leads to increased costs and longer processing times [35].

The system needs human touch, which means that individuals must verify error reports and it requires data engineers to interact with sources through multiple communication stages. All this operations results in delayed delivery of prepared data. Although automated algorithms for detecting errors have become more effective, analysts still expend a large amount of time and effort manipulating data and assessing quality issues, a process often referred to as “data wrangling” [17]. Research on immersive analytics (IA) suggests that users can perform better in pattern and anomaly detection tasks through 3D spatial interfaces rather than conventional 2D displays [14,7]. Virtual Reality (VR) environments have been successful in scientific visualization and financial modeling, reporting higher spatial interaction capabilities than spreadsheet-based reports [32].

The current state of research lacks complete understanding about how immersive technologies create new methods to verify data at its initial appearance. The current research on IA focuses on performing high-level data exploration tasks which help users understand already cleaned datasets. There is a lack of research and tools that apply VR specifically to the left side (pre-transformation stage) of the ETL process. Most ETL tools depend on 2D logs and dashboards which fail to use human spatial memory and immersive perception to detect structural data clusters and errors before they spread through the data pipeline.

In this work, we present VR-ETL, a proof-of-concept system that uses virtual reality as an alternative interface for addressing data quality issues during raw data inspection. VR-ETL allows users to inspect and correct tabular data before transformation by interacting with it inside a 3D environment. Our main contribution is the development of a spatial validation engine that maps rule-based errors (e.g., missing values, format inconsistencies) directly onto a 3D grid. However, we must explicitly note that this prototype is currently limited to CSV datasets and has been technically validated for performance with up to 10,000 records; scaling to larger datasets remains a challenge for future iterations.

To guide this exploratory study, the following research questions were defined:

- What are the latency and rendering limitations when visualizing tabular ETL data in a 3D VR environment?
- Is it technically feasible to represent rule-based validation errors spatially to improve early-stage detection?
- To what extent does a rule-based VR interface support real-time interaction for pre-transformation data correction?

The remainder of this paper is organized as follows: Section 2 provides a review of related literature on ETL technology, data quality frameworks, and immersive analytics.

Section 3 describes the system architecture and the logic of the validation engine. Section 4 details the technical implementation using Unity and OpenXR. Section 5 evaluates the results, providing an explicit discussion of benefits and limitations, while Section 6 concludes the paper and suggests directions for future research.

2. Related Work

The Extract Transform Load (ETL) operations generate the main performance problem which prevents organizations from using data for decision-making because their current systems encounter two essential problems which include extended error detection times and requirement for human verification [35,28,30]. The following section unites previous research to demonstrate VR-ETL functions as a bridge which links ETL system architectures to interactive data exploration techniques.

2.1. ETL Challenges and Real-Time Architectures

The software processing system which handles data loading and scheduled data updates operates under the name Extraction-Transformation-Loading (ETL) as it is a core function in modern data warehousing [38]. Modern ETL pipelines face their main restriction because they use reactive error management which performs validation after transformation operations which requires expensive system rollbacks [28,30,23]. The situation becomes more critical because data analysts need to perform most of their work on “data wrangling” which requires them to perform manual and repetitive data transformation and verification tasks. The main issue is that these operations remain untrackable through conventional 2D interfaces [17]. Real-time architectures achieve better output quality because their automated logging systems operate without human involvement for error correction according to Biswas et al. [6,23]. The proposed solution for these complex issues involves a direct physical Data Vault design which functions as an agile architectural method to boost ETL transformation speed and expand large-scale enterprise system capabilities [18]. The current system requires operational pre-transformation systems to run as separate units which need to identify all system irregularities at their first occurrence before human validation takes place to prevent errors from spreading through the pipeline. Moreover, other researchers have explored the use of design in enhancing ETL validation, however, only within particular domains. For instance, Saebao et al. [33] offered a method of designing ETL in business intelligence systems while considering data quality. As stated in their paper, the analysis of data quality helps make more informed decisions regarding validation steps prior to data loading. Likewise, Mhon and Kham [24] noted that when dealing with multiple sources of information, you should consider validation based on your particular scenario. The studies confirm the hypothesis that validation of input data prior to transformation is not a universal challenge that requires universal solutions, but rather a task that needs customized tools to be solved efficiently. Such a tool would be VR-ETL that introduces a validation layer in an immersive space.

2.2. Data Visualization and Business Intelligence (BI)

Existing Business Intelligence (BI) and visualization tools focus on analyzing clean post-hoc data rather than supporting interactive quality assessment of raw input [15,36]. This

focus on post-hoc analysis often ignores the fact that maintaining a high level of data quality (DQ) is a key success factor for any data warehousing system, though DQ aspects are frequently poorly defined during the initial phases of development [27]. Organizations encounter rising operational complexity because they must process multiple data types which enter their systems through various sources which complicates their ability to control DQ dimensions [5]. The current error reporting system depends on text-based output which uses static dashboards and logs that do not provide the necessary interactive features for repairing complex high-dimensional data [2,29]. Research studies about software tools for industrial real-time information processing and visualization have shown that numerous tools exist but most systems monitor performance instead of providing interactive data quality intervention capabilities [37]. The lack of interactivity is critical because identifying DQ dimensions such as completeness, accuracy, and consistency is essential for constructing a reliable and accurate analysis process [34]. As emphasized by Mause et al. [22] in ComSIS, standardized protocols for data collection and validation are necessary to evaluate data quality and minimize dataset bias. Beyond basic formatting, technical DQ also involves complex constraints like inverse referential integrity (IRICs), which are often disregarded by standard database management systems, forcing users to rely on custom, non-interactive procedures [1]. Tools such as Trifacta (previously named Wrangler) and Talend Data Preparation offer advanced two-dimensional visualization through histograms and quality bars. Unfortunately, their interface is still limited to screen visualization. On the other hand, VR-ETL intends to improve this practice through Immersive Analytics (IA). IA allows us to leave behind the dashboard visualization in favor of three-dimensional visualization where the spatial memory can be used to detect the presence of errors within multi-dimensional records.

2.3. Immersive Analytics and VR Visualization

Virtual Reality technology shows better results in pattern recognition and anomaly detection than traditional 2D dashboard systems according to Kwon et al. [19] and Bowman et al. [7]. Fonnet and Prie [14] provide a comprehensive survey of this field, tracing the evolution of immersive analytics (IA) from the early nineties to the present day and highlighting how rendering and interaction means support these complex tasks. Data specialists navigate 3D environments using analytics tools [8] and apply spatial memory to gain a more direct understanding of complex records [20,11]. The shift toward “Post-WIMP” interactions in visualization systems holds great promise for providing more natural experiences and empowering users to effectively complete analysis tasks [21]. Research studies about open-ended data exploration demonstrate that VR technology provides users with exact and detailed insights which lead to better satisfaction than conventional 2D visualization systems [25].

Research studies prove that these advantages work for particular fields which Donalek et al. demonstrate exist in astronomy [10] and trajectory data [13]. The field of collaborative visual analytics developed through the creation of Uplift which allows users to study data by using 3D models and augmented reality visualization tools for performing joint data analysis [12]. The integration of embedded data representations connects the data to the context that the data refers to. People can see clearly how the data matches the world when the integration of embedded data representations is in place [40]. However,

these systems focus on exploratory analysis of analytical results [3,4,39] rather than the operational cleaning of raw data during the ETL loading phase.

2.4. Summary of Research Gaps

As demonstrated in Table 1, VR-ETL addresses three critical gaps identified in the literature:

1. Temporal Gap: Shifting from reactive post-transformation logs to proactive pre-transformation validation [28].
2. Interaction Gap: Moving from script-based configuration to direct spatial manipulation in VR [30].
3. Application Gap: Applying immersive analytics specifically to data engineering tasks rather than final data analysis [3,20].

Table 1. Synthesis of Research Gaps in existing ETL and Immersive Analytics literature

Study	Key Contribution	Gaps Identified	Focus Area
Nwokeji & Matovu (2021)	Systematic review of ETL challenges	Focus remains on post-transformation error handling and manual corrections	ETL Error Handling
Patel & Patel (2020)	Evolution analysis of modern ETL tools	Lack of interactive, human-in-the-loop error correction mechanisms	ETL Error Handling
Simitsis et al. (2023)	Comprehensive history and future of ETL technology	Persistent challenges in real-time validation and user feedback loops	ETL Error Handling
Mehmood & Anees (2022)	Distributed real-time ETL architecture	Absence of interactive real-time correction within the stream	Real-Time ETL
Biswas et al. (2019)	Incremental loading for large-scale data integration	No VR-enhanced validation; limited to traditional 2D script monitoring	Real-Time ETL
Souibgui et al. (2020)	BI-integrated ETL systems	Primarily assumes “clean” input data; lacks error visualization	Data Vis. & BI
Geetha et al. (2020)	Evaluation of ETL tools for BI analysis	Limited interactive capabilities for handling edge-case anomalies	Data Vis. & BI
Attar-Khorasani & Chalmeta (2022)	IoT data visualization frameworks for BI	No application of visualization for ETL-stage validation	Data Vis. & BI
Pandey (2022)	Comparative study of diverse visualization tools	Does not address VR-based data engineering or ETL visualization	Data Vis. & BI
Kwon et al. (2016)	VR for financial data anomaly detection	Domain-specific application; not integrated into ETL pre-processing	VR Visualization
Bowman et al. (2012)	Theoretical benefits of 3D spatial data representation	General spatial data focus; not applied to data quality tasks	VR Visualization
Donalek et al. (2014)	Collaborative VR visualization for astronomy	High-complexity scientific visualization, not general data validation	VR Visualization
Bach et al. (2018)	VR visualization survey and research agenda	Agenda focuses on exploratory analysis, not operational validation	Immersive Analytics
Batch et al. (2020)	Expert performance metrics in immersive analytics	Focuses on analytical insight tasks rather than data cleaning	Immersive Analytics
Lee et al. (2021)	Data visceralization techniques for deep understanding	Operates under a “clean data” assumption for sensory experience	Immersive Analytics
Cordeil et al. (2017)	IATK: Immersive Analytics Toolkit development	Optimized for visualization and analysis, not data preprocessing	Immersive Analytics
Wagner Filho et al. (2020)	Evaluation of immersive scatterplot visualizations	Exploratory analysis focus; lacks record level manipulation tools	Immersive Analytics

Table 2. Comparison of Research Domains: VR Visualization vs. Traditional ETL vs. the VR-ETL Approach

Approach	Focus	Interaction Paradigm	Error Det. Timing	Visual Feedback	Primary Users	Empirical Evidence
VR Visualization Research	Analysis of clean data	Exploration and filtering	Not addressed	Analytical insights	Data analysts, researchers	Extensive user studies available
Traditional ETL Tools	Data transformation	Script based configuration	Post transformation	Log files / text dashboards	Data engineers, ETL developers	Performance benchmark available
VR-ETL (This work)	Data validation and correction	Direct manipulation in VR	Pre transformation	3D spatial visualization	Data engineers, analysts	Prototype demonstration (future studies)

The data presented in Table 2 demonstrates that VR-ETL stands alone because it uses VR visualization methods to perform ETL domain data quality operations. The research field has not studied this specific point of connection between these two areas. The prototype proves that this combination works as a technical solution which will enable researchers to study its practical benefits for ETL workflow optimization in upcoming studies.

3. VR-ETL System Architecture

3.1. Architectural Overview

The currently developed version of the VR-ETL application represents a proof-of-concept prototype system. In order to get a high-fidelity baseline for software architecture, first measurements were performed within the Unity Editor on an Apple Mac Studio machine. It should be noted that measurement within the desktop-based platform cannot completely consider the heat and performance limitations of standalone VR device (such as Meta Quest). Nevertheless, the use of the Unity Editor enabled precise measurements of the rendering pipeline and served as an inevitable intermediate stage. The VR-ETL framework is designed as a modular pipeline that transposes traditional data engineering into a 3D environment. Unlike reactive post-transformation workflows, this architecture prioritizes pre-transformation validation, allowing for human-in-the-loop inspection before data enters the transformation stage. The system follows a decoupled design consisting of five core modules: Ingestion, Validation, Rendering, Visualization, and Correction. Figure 1 illustrates the high-level architecture and data flow of the VR-ETL system.

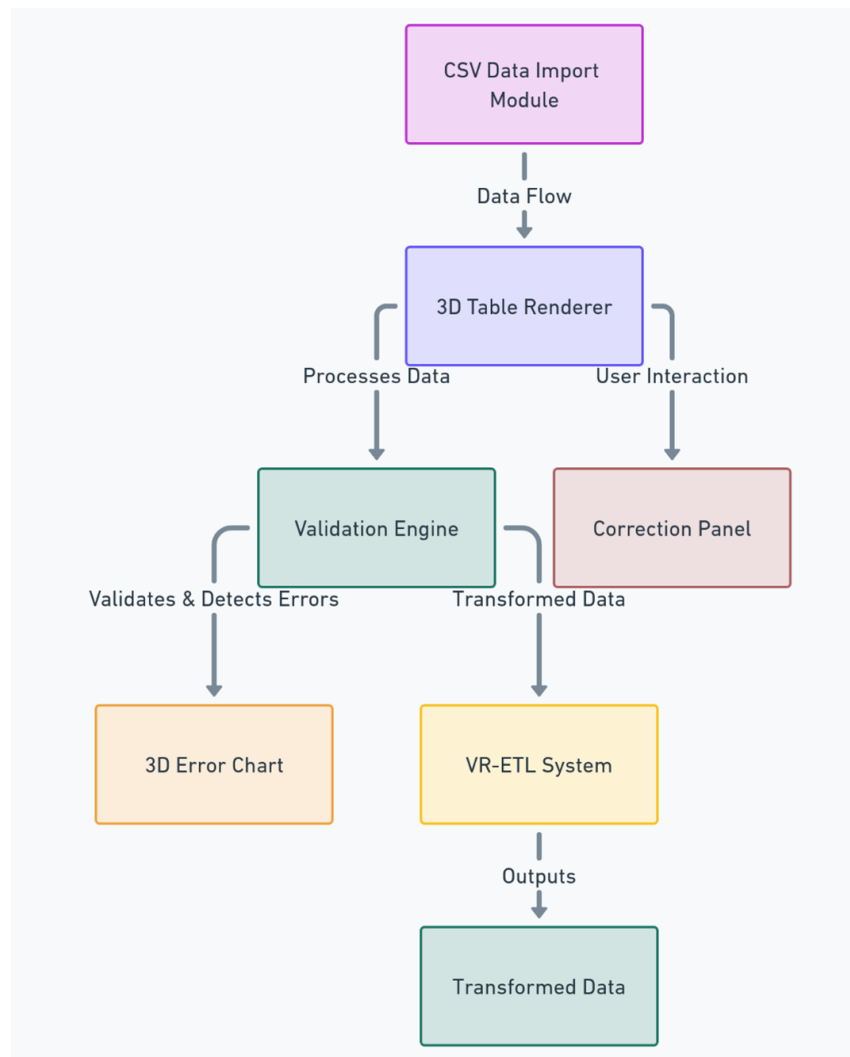


Fig. 1. Diagram of the Modular VR-ETL Pipeline showing data flow from ingestion to export

3.2. Core System Components

The system includes particular modules which transform table-based data into spatial data structures through their operations.

The CSV Ingestion Module starts by parsing the data while it detects encoding types such as UTF-8 and ASCII and determines the data types. An organized memory-based structure is created which functions as the fundamental data source for the entire pipeline process.

Validation Engine has an important role, it operates as a rule-based deterministic system that performs parallelized checks for missing data and incorrect GUIDs and date format errors and database reference errors. Error Matrix is then generated, which serves as the basis for visual feedback display.

The 3D Table Renderer module operates through Unity engine and OpenXR to convert CSV rows into particular 3D coordinates. Object Pooling is used together with View Frustum Culling to achieve the 90 FPS performance which VR requires for comfortable use.

The data is presented through two visualization methods which include the 3D Error Chart that uses spatial bar charts to analyze frequency data and the Correction Panel that shows details about single repair operations. Users can modify records through the panel which activates an automatic system to validate all changes.

3.3. Benefits and Limitations.

The VR-ETL system presents an innovative data engineering approach but exists at present as a proof-of-concept which requires definition of its benefits together with its built-in limitations.

The main advantage of VR-ETL comes from its ability to transform defensive error responses into useful predictive signals. The system enables pre-transformation validation which minimizes the expenses that occur when ETL rollbacks need to happen during late stages [28]. Users can detect error patterns through their spatial memory and peripheral vision in the 3D immersive environment which outperforms reading text-based logs. The system enables human involvement through its “human-in-the-loop” functionality which enables users to detect complex dataset structural mistakes that automated scripts fail to identify correctly.

The research contains multiple restrictions which need direct acknowledgment. The current prototype achieves its best performance when working with datasets which have records up to 10,000 in number. The frame rate in Unity becomes affected by rendering overhead when the system exceeds this threshold which could result in VR user motion sickness and reduced performance efficiency. The system operates with CSV file ingestion but it does not support processing of semi-structured data types including JSON and XML or unstructured information sources. The validation engine operates through deterministic rule-based logic which does not include machine learning or probabilistic models for anomaly detection at present. The need for specific VR equipment known as Head-Mounted Displays together with user exhaustion that occurs when using the technology for long periods create obstacles which prevent most institutions from adopting this technology.

3.4. Workflow Execution

The VR-ETL workflow follows an iterative cycle: Import → Automated Validation → Spatial Inspection → Interactive Correction → Export. By integrating the validation engine directly with the renderer, the system provides real-time visual feedback: cells transition from red (invalid) to green (corrected) instantly upon user input, eliminating the need for batch re-runs typical in tools like Talend.



Fig. 2. Interactive User Interface in the VR environment displaying real-time error detection and the correction panel

Figure 2 illustrates the workflow of the VR-ETL system from data import through validated output.

3.5. Architectural Comparison: VR-ETL vs. Traditional ETL

The design elements of VR-ETL appear in Table 3 which demonstrates how they differ from standard ETL systems through their architectural structure and implementation choices.

Table 3. Architectural and Functional Differences between Batch-based Traditional ETL and Immersive VR-ETL⁴

Feature	Traditional ETL	VR-ETL (This work)
Error Detection	Post-transformation, batch processing	Pre-transformation, real-time validation
Error Correction Method	Manual log review, SQL queries, script-based updates	Interactive VR interface with direct manipulation and immediate feedback
Error Visualization	Text logs, 2D dashboards, tabular reports	Interactive 3D error charts, spatial heatmaps, color coded cells
Data Processing Mode	Batch processing with scheduled intervals	Real-time validation during loading
Scalability	Handles millions of rows with distributed processing	Prototype tested up to 10,000 rows; larger datasets require optimization
Hardware Requirements	Standard server hardware	VR headset (Meta Quest, HTC Vive, etc.), GPU-capable PC
User Interface	Command line, GUI config panels, script editors	Immersive 3D VR environment with spatial interaction
Learning Curve	Requires SQL/scripting knowledge, ETL tool expertise	Requires VR familiarity, comparative ease untested; requires user study
Collaboration	Single user with version control, email based coordination	Prototype is single user, multi user VR planned as future work
Integration	Extensive connectors to databases, APIs, cloud services	Prototype supports CSV import/export only, integration planned
Validation Rules	Predefined and custom rules, often requires coding	Five core validation rules implemented, extensible framework
Cost	Varies (free open source to enterprise licences)	Prototype development plus VR hardware (approximately 300–500 EUR)
Maturity	Production-ready tools with extensive testing	Proof of concept prototype, not production ready

This comparison highlights important architectural differences between VR-ETL and traditional ETL approaches. VR-ETL's primary distinction is its support for pre-transformation validation with immersive interaction. However, traditional ETL tools offer advantages in scalability, ecosystem integration, and production maturity. The prototype establishes that VR-based ETL validation is technically usable for medium-sized datasets (up to 10,000 rows). Future work will determine through empirical studies whether the interactive VR approach provides measurable benefits that justify the additional hardware requirements and learning investment.

⁴ This comparison is based on the conceptual architectural design of VR-ETL versus established industry standards; empirical benchmarking between these platforms is reserved for future research.

4. Implementation & Technical Details

4.1. Overview of the VR-ETL System

The VR-ETL prototype was developed using Unity 6 and the OpenXR framework to ensure cross-platform compatibility across Meta Quest and HTC Vive ecosystems. The system architecture leverages C# for logic and Unity's World Space UI Canvas system for high-performance data rendering.

4.2. Data Ingestion

The ingestion module performs structured CSV loading through a multi-stage pipeline with three phases. First phase is an automated detection, which is used to identify UTF-8 and UTF-16 and ASCII through Byte Order Mark (BOM) analysis. Then the usage of "fail-soft" method which logs invalid rows to an audit trail while it continues to process the remaining data for uninterrupted workflow operation. Lastly, system requires 3–7 seconds to process a 10,000-row dataset on Apple M2 Max test hardware until it finishes rendering the data. To actually perform CSV parsing and type inference operations 400–600 ms is needed.

4.3. Comparison with Industry Standards: Talend and EasyMorph

The positioning of VR-ETL requires a comparison between its architectural design and the established data engineering tools Talend and EasyMorph.

The system operates through Talend which functions as its main enterprise solution for executing large-scale batch operations and providing extensive system connection abilities [15]. The main approach for error handling in Talend operates through reactive log-based systems. Engineers need to stop their main work when transformation failures occur because of poor data quality so they can review static reports and perform manual changes to either the source data or the transformation rules.

The visual interface of EasyMorph enables users to perform ETL operations through a non-technical interface which makes data transformation accessible to users who do not know how to program [29]. The system enables users to perform interactive data preparation but its 2D tabular interface on standard monitors prevents them from detecting complex multi-dimensional data relationships.

Table 4. Feature-based Comparison of the VR-ETL Prototype with Industry-Standard Tools (Talend and EasyMorph)

Feature	Talend	EasyMorph	VR-ETL (Prototype)
Primary Interface	2D GUI / Scripting	2D Visual No-code	3D Immersive VR
Validation Timing	Post-transformation	Interactive / Real-time	Pre-transformation
Error Feedback	Text-based Logs	2D Tables	3D Spatial Visuals
Correction Method	Script/Batch Update	UI-driven	Direct Spatial Manipulation

The VR-ETL system implements a pre-transformation validation mechanism which moves error detection to the initial stage of the data processing sequence. Talend provides better performance for handling large datasets but VR-ETL stands out because it enables users to perform anomaly repair through direct spatial context access. Organizations can achieve better data readiness through VR-ETL technology implementation because it creates a direct path from data reception to complete data preparation operations.

4.4. Rendering Optimizations

To maintain the 90 FPS target for VR stability, the 3D Table Renderer avoids the overhead of individual 3D meshes. Instead, it uses Unity UI Batching, which reduces draw calls by grouping text elements.

- Object Pooling: Row templates are instantiated once and recycled during scrolling to minimize Garbage Collection (GC) spikes.
- Viewport Culling: Only cells within the user’s field of view remain active, significantly reducing the GPU vertex count. In editor-based testing, this architecture maintained frame rates exceeding 500 FPS for datasets with 1,000+ visible cells.

4.5. Algorithmic Validation Logic

The Validation Engine executes a deterministic four-stage process: Type Inference, Rule Application, Flagging, and Audit Logging. The five core rule categories are implemented as follows:

- Pattern Matching: Regex-based validation for GUID and Date formats (e.g., ISO 8601).
- Uniqueness & Integrity: HashSet-based lookups provide $O(1)$ average-case complexity for duplicate primary key detection and foreign key referential integrity checks. Total validation time for 10,000 records is approximately 800 ms, ensuring that the interface remains responsive during the initial “on-ingest” validation phase.

4.6. Human-in-the-Loop Correction Workflow

The Correction Panel allows automated detection systems to operate with human judgment through its Autosuggestion System.

- Context-Aware Heuristics: For format errors (e.g., Date/GUID), the system suggests re-formatted versions based on dominant column patterns.
- Session-Based Learning: The system caches accepted user corrections to suggest consistent fixes for recurring anomalies.
- Statistical Inference: The system presents three methods to handle outliers and missing data which involve using the mean or median or choosing the most frequent value from each column’s statistical distribution. The system performs an immediate re-validation process after every correction which updates the spatial visual encoding (color change from red to green) to show instant feedback to the user.

5. Prototype Demonstration and Technical Validation

This section presents functional testing results that demonstrate the prototype's technical capabilities. These results establish feasibility but do not constitute complete evaluation, which remains as future work.

5.1. Preliminary Expert Walkthrough

The technical performance metrics received support from a first functional system evaluation which three senior data engineers performed to check the system operational sequence. The experts tested how well the VR-ETL prototype displayed tabular information through 3D visualization and they conducted tests on the process which required human involvement for data correction.

The expert walkthrough took place using a qualitative framework. Three data engineers (each having a minimum of 8 years of experience in the field) carried out a series of tasks: detection of null fields, correction of incorrect dates, and duplication of keys. The process was followed by a semi-structured interview. Though the current study is limited in terms of scale, its purpose is not to carry out any statistical analysis but to establish the grounds for future research based on the results of this qualitative study. For this reason, the NASA-TLX evaluation of workload is scheduled for a later stage of the experiment.

The participants agreed that the spatial error mapping system which used color-coded 3D tables helped them detect pattern-based anomalies better than traditional text-based error reports. This aligns with the results of Millais et al. [25], who said that while VR might not always reduce objective workload, it leads to a greater depth of insight and higher user satisfaction compared to 2D environments. The system includes real-time revalidation which enables cells to switch from red to green immediately to solve the delay problems that occur in conventional reactive ETL systems. By using these “embedded data representations” [40], the prototype leverages human spatial memory more effectively than traditional 2D wrangling interfaces described by Kandel et al. [17]. The expert feedback validated that the modular system lets users execute core data engineering tasks inside an interactive environment which will function as a base for upcoming statistical investigations.

5.2. Methodology and Test Configuration

Testing was conducted on an Apple Mac Studio (M2 Max, 32 GB RAM) using Unity 6. While functional testing included a Meta Quest 2, all quantitative performance metrics were collected within the Unity Editor to use high-fidelity profiling tools.

Three synthetic datasets were seeded with a 10% error rate across five categories (Missing Values, Malformed GUIDs, Date Errors, Duplicate Keys, and Foreign Key Violations) to establish a ground truth for validation accuracy (see Table 5).

Table 5. Ingestion and Processing Latency Metrics for datasets ranging from 50 to 10,000 records

Dataset	Rows	File Size	Total Processing Time	Scaling Factor
Small	50	2 KB	36 ms	1.0x
Medium	100	17 KB	70 ms	1.9x
Large	10,000	1.7 MB	3,451 ms	49.3x

5.3. Processing and Rendering Results

For the 10,000-row dataset, the ~ 3.5 s pipeline includes 550 ms for ingestion/parsing, 800 ms for validation, and 2.1 s for 3D instantiation. The observed sub-linear scaling (a 100x increase in rows resulting in only a 49x increase in time) suggests the architecture is efficient for medium-scale datasets.

Table 6. Success Rates of the Rule-based Engine in detecting seeded data quality anomalies

Error Type	Seeded	Detected	Rate
Pattern-Based (GUID/Date/Null)	842	842	100%
Integrity-Based (Keys/Duplicates)	93	93	100%
Total	935	935	100%

As shown in Table 6, the deterministic engine achieved a 100% detection rate for seeded rule-based errors. The Editor's system achieved more than 500 FPS (Frame Time < 2 ms) which provided sufficient resources to support the stereoscopic overhead that physical VR hardware needs. These performance metrics fulfill the high-frequency update requirements necessary for seamless immersive analytics [14]. The system maintained cell selection and feedback response times under 10 ms which fulfilled the requirements for interactive virtual reality systems and aligns with the post-WIMP interaction standards suggested by Lee et al. [21].

5.4. Comparison with Existing Literature and Limitations

While the results confirm the system's foundational logic, it is essential to position VR-ETL within the current research landscape. Unlike many immersive analytics systems that focus on the exploratory analysis of already cleaned, post-hoc analytical results [3,4], VR-ETL is specifically positioned in the operational pre-transformation phase of the ETL pipeline.

As for scalability, although 10,000 rows of data are not enough for large companies in terms of their Big Data scenarios, it was purposely selected in order to detect what part of the process is the computation cost incurred when instantiating objects in 3D space. It turned out that the 3D visualization process is the bottleneck of our system, not any calculations done with data.

In terms of hardware efficiency, our choice of a Head-Mounted Display (HMD) interface follows the findings of Cordeil et al. [9], who demonstrated that HMDs can offer faster completion times and greater accessibility for connectivity analysis compared to traditional, high-cost CAVE-style environments. However, several constraints define the current scope:

- Measurement Context: Performance was measured in the Unity Editor. Actual on-device performance (Meta Quest 2) will likely be lower due to stereoscopic rendering and wireless latency.
- Scalability Boundaries: Feasibility is confirmed up to 10,000 rows. While 2D wrangling tools like Wrangler [17] can handle much larger volumes through sampling, VR-ETL’s behavior with larger data volumes or irregular real-world encodings remains unverified due to the overhead of 3D object instantiation.
- Functional Scope: The prototype is limited to CSV files and a single-user environment, whereas systems like *Uplift* [12] have begun exploring multi-user, collaborative visual analytics.
- Absence of User Metrics: This evaluation establishes that the system works, but not that it is better than traditional 2D tools. User fatigue, the learning curve for non-VR users, and comparative efficiency against tools like Talend remain “known unknowns.”

5.5. Architectural Advantage over Traditional ETL Workflows

Although it was still in our plans to conduct empirical comparisons with other existing solutions like Talend and EasyMorph, it is possible to characterize the advantages of VR-ETL from the architecture perspective. For instance, in case of conventional systems with reactive approach, each mistake found during the transformation process requires a certain sequence of actions – checking of logs in text form, identifying the erroneous row in source data, correcting the source file itself or modification of the transformation rule and finally running the entire batch job anew. This loop may require several minutes per one run depending on the amount of data processed and usually involves context switching between at least three different applications (logs viewer, source data editor and ETL orchestration tool). VR-ETL on the contrary, enables users to fix mistakes instantly without leaving the current scene where a particular mistake was detected. The entire cycle in VR-ETL is reduced to a single step which includes instant correction of data inside the application and re-validation of changes in less than 10 ms with immediate feedback (color change from red to green provided inside VR environment). Taking into account the achieved detection rate (100% accuracy for rule-based anomalies in a dataset we used) and sublinear time scaling to 10,000 records, it is clear that our system is able to provide instant validation with no need to leave the current interface, which dramatically decreases the total cost of each correction action. In order to evaluate our solution in comparison with others such as Talend or EasyMorph we plan to perform direct benchmarks of time required for completion of identical tasks.

5.6. Validation of Research Objectives

The results confirm that the Modular Architecture (Contribution 1) and Validation Engine (Contribution 2) are functional. The 3D Interface (Contribution 3) and Error Visualiza-

tions (Contribution 4) successfully provide real-time feedback. Overall, Technical Feasibility (Contribution 5) is demonstrated within the Unity Editor environment, pending on-device validation for the 10k-row range, establishing a sound baseline for the next phase of the PhD research program. By bridging the gap between operational data engineering and immersive visualization, VR-ETL addresses a specific research void identified in the surveys of ETL [38] and Immersive Analytics [14] literature.

6. Conclusion and Future Work

Studies of virtual reality technology demonstrate its viability for ETL data validation, particularly during the critical period prior to transformation. Results from prototype development and testing indicate that VR technology can use deterministic validation methods together with 3D spatial representations of data for the detection of all pattern based anomalies within datasets while containing up to 10,000 records. The technical performance metrics support the conclusion that the makers of today's game engines have created products capable of carrying out the equivalent of real-time ETL data validation in excess of 500 FPS, which provides for an interaction latency that is less than 10 ms, thereby enabling seamless end-user experience.

The prototype proves that the concept of VR-ETL does work as intended. However, there are several issues associated with its operation that have to be overcome before implementing the concept commercially. First, it is crucial to carry out systematic validation of profiling on the Meta Quest 2 device to make sure that the results obtained through the profiling using an editor are consistent and to find out the overhead related to stereoscopic visualization and wireless communication. Additionally, the current version of the prototype allows for a single-user experience and processing of synthetic CSV files, while in actual practice, JSON and SQL formats and multiple participants should be supported.

Research collaboration on the VR ETL framework will establish a solid basis for upcoming empirical research studies. Initial effort will have a focus on short term performance improvements from the implementation of hand tracking technology and stress testing of the software with a dataset of over 50,000 records to assess critical failure points in the 3D data instantiation system. The second phase of the research will involve an empirical evaluation of the system through a controlled comparative user study with 60 or more users measuring task completion time and cognitive load using the NASA-TLX index. The long term architectural growth of this system will be achieved through three key components which will involve the use of machine learning based proactive anomaly detection and connectors to orchestrate production tools available for Apache Airflow and multi user environments to facilitate team based spatial data audits.

Research conducted utilizing the VR ETL system addresses an important gap in current research since it uses the immersive analytics to examine data processes that take place prior to the transformation of data. Also, VR Technology does not present any major technological roadblocks that would inhibit its use by organizations performing data quality management and operational data engineering. This operational standard for future experimental research studies will set guidelines for determining whether virtual reality has spatial reasoning functions that create measurable improvements in productivity and reduce mental workload for users.

In conclusion, VR-ETL demonstrates the technical feasibility of moving data validation to the pre-transformation stage within an immersive environment. As this work constitutes the foundational phase of a doctoral research program, several critical expansions are planned: hardware validation, which will include re-running all performance experiments on standalone HMD hardware to measure actual on-device latency and battery/thermal impact; user study, formal comparative study on a larger set of participants, utilizing NASA-TLX (Task Load Index) to quantitatively measure cognitive load and speed compared to traditional 2D ETL interfaces; big data integration, support for larger data sets exceeding 50,000+ records.

References

1. Aleksic, S., Ristic, S., Lukovic, I., Celikovic, M.: A design specification and a server implementation of the inverse referential integrity constraints. *Computer Science and Information Systems* 10(1), 283–320 (2013)
2. Attar-Khorasani, S., Chalmers, R.: Internet of things data visualization for business intelligence. *Big Data* 10(2), 145–165 (2022)
3. Bach, B., Sicut, R., Beyer, J., Cordeil, M., Pfister, H.: The hologram in my hand: How effective is interactive exploration of 3D visualizations in immersive tangible augmented reality? *IEEE Transactions on Visualization and Computer Graphics* 24(1), 457–467 (2018)
4. Batch, A., Cunningham, A., Cordeil, M., Elmqvist, N., Dwyer, T., Thomas, B.H., Marriott, K.: There is no spoon: Evaluating performance, space use, and presence with expert domain users in immersive analytics. *IEEE Transactions on Visualization and Computer Graphics* 26(1), 536–546 (2020)
5. Batini, C., Rula, A., Scannapieco, M., Viscusi, G.: From data quality to big data quality. In: *Big Data*, pp. 1934–1956. IGI Global (2016)
6. Biswas, N., Sarkar, A., Mondal, K.C.: Efficient incremental loading in ETL processing for real-time data integration. *International Journal of Database Theory and Application* 12(6), 1–16 (2019)
7. Bowman, B., Elmqvist, N., Jankun-Kelly, T.J.: Toward visualization for games: Theory, design space, and patterns. *IEEE Transactions on Visualization and Computer Graphics* 18(4), 662–671 (2012)
8. Cordeil, M., Cunningham, A., Dwyer, T., Thomas, B.H., Marriott, K.: Imaxes. In: *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*. pp. 71–83 (2017)
9. Cordeil, M., Dwyer, T., Klein, K., Laha, B., Marriott, K., Thomas, B.H.: Immersive collaborative analysis of network connectivity: CAVE-style or head-mounted display? *IEEE Transactions on Visualization and Computer Graphics* 23(1), 441–450 (2017)
10. Donalek, C., Djorgovski, S.G., Cioc, A., Wang, A., Zhang, J., Lawler, E., Yeh, S., Mahabal, A., Graham, M., Drake, A., Davidoff, S., Norris, J.S., Longo, G.: Immersive and collaborative data visualization using virtual reality platforms. In: *2014 IEEE International Conference on Big Data (Big Data)*. pp. 609–614 (2014)
11. Dwyer, T., Marriott, K., Isenberg, T., Klein, K., Riche, N., Schreiber, F., Stuerzlinger, W., Thomas, B.H.: *Immersive Analytics: An Introduction*. KOPS Universität Konstanz (2018)
12. Ens, B., Goodwin, S., Prouzeau, A., Anderson, F., Wang, F.Y., Gratzl, S., Lucarelli, Z., Moyle, B., Smiley, J., Dwyer, T.: Uplift: A tangible and immersive tabletop system for casual collaborative visual analytics. *IEEE Transactions on Visualization and Computer Graphics* 27(2), 1193–1203 (2021)

13. Filho, J.A., Stuerzlinger, W., Nedel, L.: Evaluating an immersive space-time cube geovisualization for intuitive trajectory data exploration. *IEEE Transactions on Visualization and Computer Graphics* 26(1), 514–524 (2020)
14. Fonnet, A., Prie, Y.: Survey of immersive analytics. *IEEE Transactions on Visualization and Computer Graphics* 27(3), 2101–2122 (2021)
15. Geetha, S., Dhanani, K.R., Doshi, P.P.: Data analysis and ETL tools in business intelligence. *International Journal of Multidisciplinary and Current Research* 8(1), 58–64 (2020)
16. Gupta, K., Jiwani, N.: A systematic overview of fundamentals and methods of business intelligence. *International Journal of Science and Digital Convergence* 7(2), 145–151 (2021)
17. Kandel, S., Paepcke, A., Hellerstein, J., Heer, J.: Wrangler. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. pp. 3363–3372 (2011)
18. Krneta, D., Jovanovic, V., Marjanovic, Z.: A direct approach to physical data vault design. *Computer Science and Information Systems* 11(2), 569–599 (2014)
19. Kwon, O.H., Muelder, C., Lee, K., Ma, K.L.: A study of layout, rendering, and interaction methods for immersive graph visualization. *IEEE Transactions on Visualization and Computer Graphics* 22(1), 680–689 (2016)
20. Lee, B., Brown, D., Lee, B., Hurter, C., Drucker, S., Dwyer, T.: Data visceralization: Enabling deeper understanding of data using virtual reality. *IEEE Transactions on Visualization and Computer Graphics* 27(2), 1095–1105 (2021)
21. Lee, B., Srinivasan, A., Isenberg, P., Stasko, J.: Post-WIMP interaction for information visualization. *Foundations and Trends in Human–Computer Interaction* 14(1), 1–95 (2021)
22. Mause, G., Galinac-Grbac, T., Dalbelo-Basic, B.: A systematic data collection procedure for software defect prediction. *Computer Science and Information Systems* 13(1), 173–197 (2016)
23. Mehmood, E., Anees, T.: Distributed real-time ETL architecture for unstructured big data. *The Journal of Supercomputing* 78(8), 10803–10832 (2022)
24. Mhon, G.G.W., Kham, N.S.M.: ETL preprocessing with multiple data sources for academic data analysis. In: *2020 16th International Conference on Computing, Communication and Networking Technologies (ICCN)*. pp. 1–5 (2020)
25. Millais, P., Jones, S.L., Kelly, R.: Exploring data in virtual reality. In: *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. pp. 1–6 (2018)
26. Mondal, K.C., Biswas, N., Saha, S.: Role of machine learning in ETL automation. In: *Proceedings of the 3rd International Conference on Vision, Image and Signal Processing*. pp. 1–5 (2020)
27. Nemani, R.R., Konda, R.: A framework for data quality in data warehousing. In: *Lecture Notes in Business Information Processing*, pp. 292–297. Springer (2009)
28. Nwokeji, J.C., Matovu, R.: A systematic literature review on big data extraction, transformation and loading (ETL). In: *Advances in Intelligent Systems and Computing*, vol. 1249, pp. 259–271. Springer (2021)
29. Pandey, A.: Comparative study of data visualization tools in big data analysis for business intelligence. *International Journal for Research in Applied Science and Engineering Technology* 10(6), 2591–2600 (2022)
30. Patel, M., Patel, D.B.: Progressive growth of ETL tools: A literature review of past to equip future. In: *Advances in Intelligent Systems and Computing*, vol. 1186, pp. 501–510. Springer (2020)
31. Qaiser, A., Farooq, M.U., Mustafa, S.M.N., Abrar, N.: Comparative analysis of ETL tools in big data analytics. *Pakistan Journal of Engineering and Technology* 1(1), 1–8 (2023)
32. Ragan, E.D., Sowndarajan, A., Kopper, R., Bowman, D.A.: The effects of higher levels of immersion on procedure memorization performance and implications for educational virtual environments. *IEEE Transactions on Visualization and Computer Graphics* 16(5), 891–901 (2010)

33. Saebao, S., Matayong, S., Trakulmaykee, N.: QoX based ETL design for business intelligence system of lecturers' qualifications analysis. In: 2020 17th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON). pp. 465–468 (2020)
34. Sidi, F., Shariat Panahy, P.H., Affendey, L.S., Jabar, M.A., Ibrahim, H., Mustapha, A.: Data quality: A survey of data quality dimensions. In: 2012 International Conference on Information Retrieval & Knowledge Management. pp. 300–304 (2012)
35. Simitsis, A., Skiadopoulos, S., Vassiliadis, P.: The history, present, and future of ETL technology. *CEUR Workshop Proceedings* 3351, 1–12 (2023)
36. Souibgui, M., Atigui, F., Ben Yahia, S., Si-Said Cherfi, S.: Business intelligence and analytics: On-demand ETL over document stores. In: *Communications in Computer and Information Science*, vol. 1221, pp. 659–673. Springer (2020)
37. Sousa, R., Miranda, R., Moreira, A., Alves, C., Lori, N., Machado, J.: Software tools for conducting real-time information processing and visualization in industry: An up-to-date review. *Sustainability* 11(11), 4800 (2021)
38. Vassiliadis, P.: A survey of extract–transform–load technology. *International Journal of Data Warehousing and Mining* 5(3), 1–27 (2009)
39. Wagner Filho, J.A., Rey, M.F., Freitas, C.M., Nedel, L.: Immersive visualization of abstract information: An evaluation on dimensionally-reduced data scatterplots. In: 2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR). pp. 483–490 (2020)
40. Willett, W., Jansen, Y., Dragicevic, P.: Embedded data representations. *IEEE Transactions on Visualization and Computer Graphics* 23(1), 461–470 (2017)

Maja Cerjan is a teaching assistant at the Faculty of Organization and Informatics (FOI), University of Zagreb, Croatia, within the Department of Theoretical and Applied Foundations of Information Sciences, where she teaches courses related to databases, data warehouses, and business intelligence. She is a final-year doctoral candidate in the Information Society programme at the Faculty of Information Studies (FIŠ) in Novo Mesto, Slovenia. Her doctoral research focuses on the application of virtual reality to the design and validation of ETL (Extract, Transform, Load) processes. Her research interests include ETL processes, data warehousing, dimensional modelling, business intelligence, deductive databases, and the use of immersive technologies in data engineering.

Kornelije Rabuzin is a Full Professor at the Faculty of Organization and Informatics (FOI), University of Zagreb, Croatia, where he heads the Laboratory for Data Technologies within the Department of Theoretical and Applied Foundations of Information Sciences. He teaches several courses related to databases and data warehouses. He is the author and co-author of nearly 100 scientific papers and around 10 books and book chapters in the fields of databases, data warehouses, and business intelligence. His research interests include databases, data warehouses, business intelligence, big data, and, in recent years, graph databases and graph data warehouses, including integrity constraints and query languages for graph databases.

Leo Mršić received his PhD in information sciences from the Faculty of Humanities and Social Sciences, University of Zagreb, Croatia, in 2012, with research focused on graph theory and the application of statistical models in business. He is a Full Professor active in the scientific fields of computing, information sciences, and economics, and

currently serves as Vice-Rector for Science and Research at Algebra Bernays University in Zagreb. He is the author and co-author of more than 100 scientific papers and holds several patents. He is an IEEE Senior Member, an ACM member, and an IPMA Level A certified project director. His research interests include data science, business intelligence, data management and analysis, and artificial intelligence.

Received: February 19, 2026; Accepted: July 15, 2026.

