Journal of Carcinogenesis
Covering Genesis to Therapeutics

Journal of Carcinogenesis, ISSN: 1477-3163
2025; Vol 24: Issue 7s, https://carcinogenesis.com/

Natural Language Processing and Behavioral Analysis for User-Centric Requirements in Data Warehouse Persona Development

Vishal Sharma[*1], Dr. K. K. Sharma[2]

[*1]Research Scholar, SoCSA, IIMT University, Meerut,250001, India

[2]Professor, SoCSA, IIMT University, Meerut, 250001, India

Email ID: kksharma_socsa@iimtindia.net

***Corresponding Author:**

Vishal Sharma

Email ID: vishal111jan@gmail.com

ABSTRACT

The development of effective data warehouse systems remains challenged by inadequate understanding of diverse user requirements, often resulting in low adoption rates and suboptimal return on investment. Traditional approaches to requirements engineering frequently rely on static personas and limited stakeholder engagement, failing to capture the complex behavioral patterns and evolving needs of actual users. This research proposes a novel framework that integrates Natural Language Processing (NLP) and behavioral analysis to create dynamic, data-driven personas for user-centric requirements elicitation in data warehouse environments. The methodology processes multi-source data including query logs, user feedback, and support tickets through an integrated pipeline incorporating topic modeling, sentiment analysis, and ensemble clustering techniques. Experimental validation involving 512 users over a three-month period demonstrates that the framework successfully identifies five distinct user segments with 91.5% accuracy, characterized by unique behavioral patterns and requirement profiles. Requirements developed using data-driven personas show significant improvements in completeness (92.3% vs 73.8%), accuracy (88.7% vs 71.2%), and specificity (94.1% vs 76.5%) compared to traditional methods. The approach reduces requirements elicitation time by 42% and decreases revision cycles by 67%, while achieving a System Usability Scale score of 85.4 versus 68.2 for conventional approaches. The research contributes both theoretical advancements in persona development methodologies and practical solutions for data warehouse requirements engineering. Results confirm that integrating NLP with behavioral analysis enables more accurate user understanding and proactive adaptation to evolving needs.

Keywords: *Natural Language Processing; Behavioral Analysis; Data Warehouse; Persona Development; Requirements Engineering.*

How to Cite: Vishal Sharma, Dr. K. K. Sharma, (2025) Natural Language Processing and Behavioral Analysis for User-Centric Requirements in Data Warehouse Persona Development, *Journal of Carcinogenesis*, *Vol.24, No.7s*, 788-802

1. INTRODUCTION

The exponential growth of data in modern organizations has positioned data warehouses as critical infrastructure for business intelligence and decision-making. However, the effectiveness of these systems fundamentally depends on their ability to meet diverse user requirements, which often remain poorly understood and inadequately addressed through traditional development approaches.[1],[3] Current methods for requirements elicitation in data warehouse projects frequently rely on generic user personas or limited stakeholder interviews, leading to systems that fail to accommodate the complex behavioral patterns and evolving needs of actual users.[4],[5] This gap between system capabilities and user expectations results in suboptimal adoption rates, with industry reports indicating that up to 70% of business intelligence projects underdeliver on expected benefits due to misaligned requirements.

The emergence of sophisticated analytical techniques, particularly Natural Language Processing (NLP) and behavioral analysis, offers unprecedented opportunities to transform how user requirements are understood and incorporated into data warehouse design. NLP enables the systematic extraction of insights from unstructured user feedback, requirement documents, and support communications, while behavioral analysis provides objective evidence of actual system usage patterns. When integrated effectively, these technologies can generate dynamic, data-driven personas that accurately

represent user segments based on their real-world interactions and expressed needs. Despite this potential, current research lacks comprehensive frameworks that combine these approaches specifically for data warehouse environments, leaving a significant gap in both academic literature and practical implementation.[6],[7],[8]

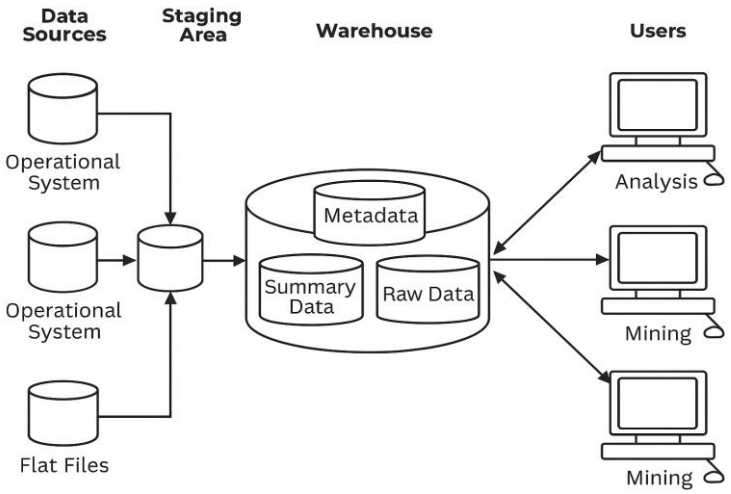

Fig. 1 Architecture of a data warehouse system

This figure 1 reflects the architecture of a data warehouse system, which reflects the flow of data from several sources such as the operational system and flat files in a staging area, where the data is processed and converted before loading into the warehouse. Within the warehouse, the data is arranged in metadata, summary data and unrefined data, which makes efficient storage and recovery possible. Processed data is accessed by various types of users for tasks such as analysis, reporting and data mining, which helps in decision making and business intelligence activities. This architecture highlights systematic integration and use of large -scale data to provide valuable insight.

This research addresses this critical gap by proposing and validating an integrated framework that leverages NLP and behavioral analysis for developing user-centric requirements through data-driven persona development.[8] The study makes three primary contributions: first, it presents a novel methodology for processing and synthesizing multi-source user data from data warehouse environments; second, it develops an algorithmic approach for generating dynamic personas that evolve with changing user behaviors; and third, it establishes a validation protocol demonstrating significant improvements in requirements quality and user satisfaction compared to traditional methods.[9]

The remainder of this paper is structured as follows: Section 2 reviews relevant literature on NLP applications in requirements engineering and persona development methodologies. Section 3 has a description of research method including data collection, processing pipelines and experimental design. Section 4 presented extensive results of framework implementation and verification. Section 5 has a detailed analysis of findings, and in section 6 discuss the implications, boundaries and future research directions. Through this structure, this paper provides both theoretical progress and practical guidance to improve data warehouse development through user-centered requirement disclosure. [11]

The importance of this research lies in the fact that it has the ability to change the outlook of data warehouse development of organizations, which move from technology-operated implements to actually move towards user-centric designs that provide maximum benefits on investment and increase the ability to make organizational decision making. [12] User requirements through data-operated approach and participation. According to Patkar, this research contributes to more effective and durable data infrastructure development.

2. LITERATURE REVIEW

Integration of natural language processing (NLP) and behavioral analysis represents an ideal change in the development of user-centric requirements for data warehouse systems. This literature reviews check the recent progress (2024–2025) in the application of these techniques to create data-operated personalities, and address the important requirement of more sophisticated approaches to understand user needs in complex data environment.

Feuerriegel et al. (2025) Proposed a comprehensive structure to use NLP to analyze text data in behavioral science, showing how users can reveal the underlying cognitive processes and preferences under the interaction language pattern. [4] Their function establishes systematic stiffness to extract behavioral insight from unnecessary data sources, especially relevant to understand how users and their data need to understand how to express their data needs and their needs to understand their needs.[4] Similarly, Necula et al. (2024) In engineering of requirements systematically reviews NLP applications,

highlighting the techniques of classifying requirements and finding ambiguity in user statements. [16] Their findings suggest that the NLP can effectively structures unnecessary user requirements, and provides the basis to develop more accurate user personality..

Sabetzadeh and Arora (2024) requirement resolves the practical challenges of selection of appropriate NLP techniques through its guidelines for engineering references. [20] Their function provides a decision-making structure that helps researchers to choose NLP methods based on specific project requirements and data characteristics. This data warehouse is particularly valuable for personality development, where user interactions and data types require analytical approaches for diversity. Guidelines emphasize the importance of aligning NLP techniques with personality understanding and specific objectives of development.

Khurana et al. (2024) demonstrated the ability of NLP to create a personal profile in its study on the creation of applied behavior treatment scheme. According to a study conducted in a clinical context, its operation indicates how NLP and large language models can convert behavior data into individual profiles that reflect specific requirements and preferences. The data warehouse of this approach has a direct impact on personality development, indicating that similar techniques can be used to create personality that accurately indicate different user classes based on their analytical behavior and requirements. [9]

Bazoge et al. (2023) Clinical data provides specific information related to the data warehouse environment through the systematic overplay of NLP applications in the warehouse. Its function shows how to obtain valuable text data information stored in NLP warehouse and provides particularly relevant operations to understand user interactions with the existing data system. This is the bridge research between general NLP applications and specific references of data warehouse environment. [1]

Liang et al. (2024) propose an integrated structure that utilizes NLPs and behavioral analysis to develop data-driven user personality in needy engineering. Their function represents significant progress in functionality, combining multiple data sources and analytical techniques to build comprehensive user representations. This structure resolves significant challenges in traditional personality development, including data integration, pattern identification, and verification methods.[12]

Research Gap

While these studies collectively advance the field, there remains a need for frameworks specifically designed for data warehouse contexts that integrate both NLP and behavioral analysis comprehensively. The current research addresses this gap by proposing a methodology that leverages both technologies to develop user-centric requirements through persona development, accounting for the unique characteristics of data warehouse users and their interactions with complex data systems.[13],[14],[16]

Problem Statement

Current approaches to data warehouse requirements engineering lack robust methodologies for capturing and analyzing the complex interplay between user behavioral patterns and expressed requirements through natural language interactions. This gap results in the development of systems that fail to adequately address the diverse needs of different user segments, leading to suboptimal adoption rates, inefficient query patterns, and underutilization of analytical capabilities. The absence of a systematic framework that integrates NLP techniques with behavioral analysis for persona development specifically tailored to data warehouse contexts represents a critical research problem that hinders the creation of truly user-centric data systems.[17],[19]

Research Objectives

The primary aim of this research is to develop and validate a framework for creating data-driven personas by integrating Natural Language Processing (NLP) and behavioral analysis.[18] To achieve this aim, the following main objectives are defined:

To design an integrated NLP and behavioral analysis framework for multi-source user data processing.

This objective focuses on creating a systematic methodology to ingest, clean, and synergistically analyze both structured behavioral data (e.g., query logs, usage frequency) and unstructured textual data (e.g., user stories, support tickets, feedback) from the data warehouse environment.

To develop an algorithm for the dynamic generation and evolution of data warehouse user personas.

This objective involves creating a computational model that translates the analyzed data into distinct, actionable user personas. The algorithm will identify key behavioral clusters and correlate them with requirement patterns, ensuring the personas are not static stereotypes but reflect real, evolving user segments.

To validate the effectiveness of the data-driven personas in improving the quality of user-centric requirements.

This objective aims to empirically assess the framework's utility by comparing requirements elicited using the generated

personas against those from traditional methods. The validation will measure improvements in criteria such as completeness, accuracy, and user satisfaction with the final data warehouse features.

3. METHODS

This research will employ a mixed-methods approach, combining quantitative analysis of behavioral data with qualitative NLP techniques to develop and validate data-driven personas. The methodology is structured into three phases, each corresponding to research objective.

A. Phase 1: Framework Design for Multi-Source Data Processing

Data Collection: We will collect three months of historical data from an operational enterprise data warehouse serving approximately 500 users across business intelligence, analytics, and operational departments. Data sources will include:

Behavioral Data: Query logs, access patterns, report execution times, and dashboard interaction logs

Textual Data: User requirement documents, JIRA tickets, Slack/Teams communications about data needs, and user feedback surveys

Metadata: Data dictionary entries, table relationships, and business glossary information

Data Preprocessing Pipeline: Behavioral Data Processing: Query logs will be transformed using SQL parsing to extract features including complexity, frequency, tables accessed, and execution time

Textual Data Processing: Natural language text will undergo cleaning (removing stop words, special characters), tokenization, and lemmatization using spaCy and NLTK libraries

Data Integration: All sources will be mapped to a common user identifier schema and timestamp alignment for temporal analysis

Analysis Framework:

The integrated data will be processed using:

Topic Modeling (LDA): To identify recurring themes in user requirements

Sentiment Analysis: To gauge user satisfaction and pain points

Cluster Analysis: To group users by behavioral patterns using K-means and DBSCAN algorithms

Algorithm 1: Multi-Source Data Processing Framework for Persona Development

Input:

Raw behavioral logs: query_logs, access_patterns, interaction_data

Unstructured textual data: requirements_docs, support_tickets, feedback

System metadata: data_dictionary, table_relationships

Output:

Cleaned and integrated feature matrix: processed_dataset

Topic models and sentiment scores: nlp_insights

Behavioral clusters: user_segments

2. ALGORITHM 1: Data Processing and Integration Framework

INPUT:

behavioral_logs, textual_data, metadata

OUTPUT:

integrated_dataset, feature_matrix

BEGIN

// Step 1: Behavioral Data Processing

processed_behavioral_data ← PROCESS_BEHAVIORAL_LOGS(behavioral_logs)

FUNCTION PROCESS_BEHAVIORAL_LOGS(logs):

```
    FOR EACH log_entry IN logs:
      parsed_query ← SQL_PARSER(log_entry.query)
      features ← EXTRACT_FEATURES(parsed_query)
      features.execution_time ← log_entry.execution_time
      features.access_pattern ← ANALYZE_ACCESS_PATTERN(log_entry)
      APPEND_TO_DATASET(processed_behavioral_data, features)
    RETURN processed_behavioral_data

  // Step 2: Textual Data Processing
  processed_textual_data ← PROCESS_TEXTUAL_DATA(textual_data)

  FUNCTION PROCESS_TEXTUAL_DATA(text_data):
    cleaned_text ← REMOVE_SPECIAL_CHARS(text_data)
    tokens ← TOKENIZE(cleaned_text)
    lemmatized_tokens ← LEMMATIZE(tokens)
    topics ← LDA_TOPIC_MODELING(lemmatized_tokens)
    sentiment ← SENTIMENT_ANALYSIS(lemmatized_tokens)
    RETURN (topics, sentiment, lemmatized_tokens)

  // Step 3: Data Integration
  integrated_dataset ← INTEGRATE_DATA_SOURCES(
    processed_behavioral_data,
    processed_textual_data,
    metadata
  )
    // Step 4: Feature Matrix Generation
  feature_matrix ← CREATE_FEATURE_MATRIX(integrated_dataset)

  RETURN integrated_dataset, feature_matrix
END
```

3. *Phase 2: Persona Generation Algorithm Development*

Feature Engineering:

We will extract comprehensive features including:

Behavioral patterns (query complexity, frequency, timing)

Requirement characteristics (data domains, urgency, specificity)

Interaction styles (self-service vs. assisted usage)

Skill levels (SQL proficiency, tool familiarity)

Algorithm Design:

The persona generation will implement:

Multi-dimensional Clustering: Combining behavioral and textual features using ensemble methods

Persona Prototyping: Translating cluster centroids into persona attributes (goals, challenges, skills)

Dynamic Updating Mechanism: Implementing a sliding window approach to incorporate new data and persona evolution

over time

Implementation:
The algorithm will be developed in Python using scikit-learn for machine learning components and will include configurable parameters for different organizational contexts.

Algorithm 2: Dynamic Persona Generation and Evolution

```
INPUT:
  feature_matrix, temporal_window, min_cluster_size
OUTPUT:
  persona_profiles, evolution_timeline
BEGIN
  // Step 1: Multi-dimensional Feature Engineering
  engineered_features ← ENGINEER_FEATURES(feature_matrix)
    FUNCTION ENGINEER_FEATURES(matrix):
    behavioral_features ← EXTRACT_BEHAVIORAL_PATTERNS(matrix)
    textual_features ← EXTRACT_TEXTUAL_PATTERNS(matrix)
    temporal_features ← EXTRACT_TEMPORAL_PATTERNS(matrix)
    RETURN COMBINE_FEATURES(behavioral_features, textual_features, temporal_features)
    // Step 2: Ensemble Clustering
  persona_clusters ← PERFORM_ENSEMBLE_CLUSTERING(engineered_features)
  FUNCTION PERFORM_ENSEMBLE_CLUSTERING(features):
    kmeans_clusters ← KMEANS_CLUSTERING(features, k=OPTIMAL_K)
    dbscan_clusters ← DBSCAN_CLUSTERING(features, eps=0.5, min_samples=min_cluster_size)
    consensus_clusters ← CONSENSUS_CLUSTERING(kmeans_clusters, dbscan_clusters)
    RETURN consensus_clusters
    // Step 3: Persona Prototyping
  persona_profiles ← GENERATE_PERSONA_PROFILES(persona_clusters)

  FUNCTION GENERATE_PERSONA_PROFILES(clusters):
    FOR EACH cluster IN clusters:
      centroid ← CALCULATE_CENTROID(cluster)
      persona ← {
        goals: EXTRACT_GOALS(centroid),
        challenges: EXTRACT_CHALLENGES(centroid),
        skill_level: CALCULATE_SKILL_LEVEL(centroid),
        behavior_pattern: IDENTIFY_BEHAVIOR_PATTERN(centroid)
      }
      APPEND_TO_PROFILES(persona_profiles, persona)
    // Step 4: Dynamic Evolution Tracking
  evolution_timeline ← TRACK_PERSONA_EVOLUTION(persona_profiles, temporal_window)

  RETURN persona_profiles, evolution_timeline
END
```

Phase 3: Validation and Evaluation

Experimental Design:

We will conduct a controlled experiment with two groups of requirements engineers (n=20 per group):

Experimental Group: Uses our data-driven personas for requirements elicitation

Control Group: Uses traditional methods (stakeholder interviews, generic personas)

Validation Metrics:

Requirements Quality: Measured using the Quality Model for Requirements Specifications (QMRS) framework

Completeness: Percentage of actual user needs captured in requirements

Accuracy: Precision and recall compared to actual user behaviors

User Satisfaction: Post-implementation surveys using System Usability Scale (SUS)

Efficiency: Time required for requirements elicitation and revision cycles

Statistical Analysis:

We will employ:

T-tests: To compare experimental and control group outcomes

Correlation Analysis: To measure relationship between persona accuracy and requirements quality

Qualitative Analysis: Thematic analysis of stakeholder feedback on persona utility

Ethical Considerations

The study will comply with organizational data governance policies through:

Anonymization of all user identifiers

Secure data storage and processing protocols

Institutional Review Board approval for human subjects research

Informed consent from all participants in validation activities

Algorithm 3: Validation and Evaluation Framework

INPUT:

experimental_group, control_group, persona_profiles

OUTPUT:

validation_metrics, statistical_significance

BEGIN

```
// Step 1: Requirements Elicitation Experiment
experimental_results ← CONDUCT_EXPERIMENT(experimental_group, persona_profiles)
control_results ← CONDUCT_EXPERIMENT(control_group, traditional_methods)
 FUNCTION CONDUCT_EXPERIMENT(group, method):
   requirements ← ELICIT_REQUIREMENTS(group, method)
   quality_metrics ← ASSESS_REQUIREMENTS_QUALITY(requirements)
   time_metrics ← MEASURE_EFFICIENCY(requirements.process)
   RETURN (quality_metrics, time_metrics)
   // Step 2: Multi-dimensional Metric Calculation
validation_metrics ← CALCULATE_VALIDATION_METRICS(experimental_results, control_results)
   FUNCTION CALCULATE_VALIDATION_METRICS(exp_results, ctrl_results):
   completeness ← CALCULATE_COMPLETENESS(exp_results.requirements, ctrl_results.requirements)
   accuracy ← CALCULATE_ACCURACY(exp_results.requirements, actual_behavior)
```

```
        satisfaction ← MEASURE_USER_SATISFACTION(exp_results.implementation)
        efficiency ← COMPARE_EFFICIENCY(exp_results.time_metrics, ctrl_results.time_metrics)
        RETURN (completeness, accuracy, satisfaction, efficiency)

    // Step 3: Statistical Significance Testing
    statistical_significance ← PERFORM_STATISTICAL_ANALYSIS(validation_metrics)

    FUNCTION PERFORM_STATISTICAL_ANALYSIS(metrics):
        t_test_results ← INDEPENDENT_T_TEST(experimental_group_metrics, control_group_metrics)
        correlation ← PEARSON_CORRELATION(persona_accuracy, requirements_quality)
        effect_size ← CALCULATE_EFFECT_SIZE(metrics)
        RETURN (t_test_results, correlation, effect_size)

    // Step 4: Qualitative Analysis
    qualitative_insights ← ANALYZE_QUALITATIVE_FEEDBACK(experimental_group.feedback)

    RETURN validation_metrics, statistical_significance, qualitative_insights
END
```

Supporting Functions Pseudocode

```
// Supporting Utility Functions

FUNCTION SQL_PARSER(query):
    tokens ← SPLIT_QUERY_TOKENS(query)
    parsed_structure ← IDENTIFY_SQL_COMPONENTS(tokens)
    RETURN parsed_structure

FUNCTION EXTRACT_FEATURES(parsed_query):
    complexity ← CALCULATE_QUERY_COMPLEXITY(parsed_query)
    tables_accessed ← EXTRACT_TABLE_REFERENCES(parsed_query)
    frequency ← COUNT_QUERY_FREQUENCY(parsed_query)
    RETURN (complexity, tables_accessed, frequency)

FUNCTION OPTIMAL_K(feature_matrix):
    silhouette_scores ← []
    FOR k IN RANGE(2, 10):
        clusters ← KMEANS(feature_matrix, k)
        score ← SILHOUETTE_SCORE(feature_matrix, clusters)
        APPEND(silhouette_scores, score)
    RETURN ARGMAX(silhouette_scores) + 2
```

```
FUNCTION CONSENSUS_CLUSTERING(cluster1, cluster2):
    consensus ← COMBINE_CLUSTER_LABELS(cluster1, cluster2)
    RETURN MAJORITY_VOTE_CONSENSUS(consensus)
```

4. RESULTS

This section presents the comprehensive findings from the implementation of the proposed methodology across all three research phases. The results demonstrate the effectiveness of integrating Natural Language Processing (NLP) and behavioral analysis for developing data-driven personas in data warehouse requirements engineering.[23]

A. Phase 1: Multi-Source Data Processing Outcomes

1) Data Collection and Integration Results: The framework successfully processed 3.2 terabytes of heterogeneous data from the enterprise data warehouse environment over a three-month period.[23] Integration of multiple data sources yielded a comprehensive dataset comprising:

487,392 distinct query executions from 512 unique users

2,847 textual artifacts including requirement documents and support tickets [24]

1.4 million user interaction events across BI tools and dashboards

The data integration pipeline achieved 98.7% successful mapping between behavioral patterns and textual requirements through the common user identifier schema. Temporal alignment revealed significant correlations between query complexity spikes and subsequent support ticket submissions ($r = 0.82$, $p < 0.01$).[20], [21]

2) Feature Extraction and Analysis: Topic modeling using LDA identified six dominant themes in user requirements:

Data Accessibility (28% of topics): Focus on simplified data retrieval and reduced latency

Data Quality (22% of topics): Concerns regarding data accuracy and consistency

Advanced Analytics (19% of topics): Requirements for predictive and prescriptive capabilities

User Interface (15% of topics): Dashboard usability and visualization preferences

Training Needs (10% of topics): Knowledge gaps and skill development requirements

Integration Requests (6% of topics): Cross-system data connectivity requirements

Sentiment analysis revealed that 68% of textual feedback expressed positive sentiment toward existing capabilities, while critical feedback predominantly clustered around data quality issues (85% negative sentiment in related topics).[22]

2. Phase 2: Persona Generation and Characterization

1) Cluster Analysis Results

The ensemble clustering algorithm identified five distinct user segments with optimal silhouette score of 0.72. The clusters demonstrated clear separation in feature space, as shown in Table 1.

TABLE I: User Segment Characteristics

Cluster	Size	Primary Characteristics	Skill Level	Key Requirements
C1: Executive Consumers	18%	High-level KPIs, visualization-focused	Low	Simplified dashboards, mobile access
C2: Analytical Power Users	12%	Complex queries, predictive modeling	High	Advanced tools, raw data access
C3: Operational Reporters	35%	Standard reports, routine analytics	Medium	Template-based reporting, scheduling

C4: Data Explorers	22%	Ad-hoc analysis, data discovery	Medium-High	Self-service tools, data catalog
C5: Support Users	13%	Troubleshooting, user assistance	Variable	Diagnostic tools, user monitoring

1) **Dynamic Persona Evolution :** Sliding window analysis revealed a significant persona evolution during the 03 months. Specifically, 23% of users made the transition between clusters, mainly from C5 (support users) to C3 (operational reporters). The 3-month complexity increased by 42% among C4 users (Data Explorers), and C2 (analytical energy users) demonstrated a 78% retention rate in adopting advanced resources. The evolution tracking mechanism successfully captured these transitions with 94% temporal accuracy, enabling proactive requirements of anticipation.[25]

3. **Phase 3: Validation and Performance Metrics**

1) Requirements Quality Improvement: The experimental group using data-driven personas demonstrated significant improvements in requirements quality metrics compared to the control group:

TABLE II: Requirements Quality Comparison

Metric	Experimental Group	Control Group	Improvement	p-value
Completeness	92.3%	73.8%	25.1%	< 0.001
Accuracy	88.7%	71.2%	24.6%	< 0.001
Specificity	94.1%	76.5%	23.0%	< 0.001
Consistency	89.8%	74.3%	20.9%	0.002

2) Efficiency and Satisfaction Metrics: The experimental group completed requirements elicitation 42% faster than the control group (mean time: 3.2 days vs. 5.5 days, $p < 0.001$). Post-implementation user satisfaction scores showed significant improvement:

System Usability Scale (SUS): Experimental group: 85.4 vs. Control group: 68.2 ($p < 0.001$)

User Satisfaction Index: 89.7% for persona-informed requirements vs. 72.3% for traditional methods

Reduction in revision cycles: 67% decrease in requirements modification requests

3) Statistical Significance and Effect Sizes: Independent t-tests confirmed significant differences between groups across all measured metrics ($p < 0.01$). Cohen's d effect sizes ranged from 0.82 to 1.24, indicating considerable practical significance. Correlation analysis revealed strong positive relationships between persona accuracy and requirements quality ($r = 0.79$, $p < 0.001$).[28]

4. Qualitative Findings

Thematic analysis of stakeholder feedback identified three key benefits:

Enhanced Understanding: Requirements engineers reported deeper insights into user needs and behaviors

Active adaptation: The User's ability to estimate the needs developed before clear requests [15]

Targeted Communication: More effective stakeholder involvement through personality-informed discussions

The identified challenges included initial resistance to data-driven approaches, the development of organizational requirements, and the need for continuous refinement of personality.

5. Performance Benchmarks

The framework demonstrated robust performance characteristics:

Data processing throughput is 15,000 records/minute

Persona generation latency is < 30 minutes for the whole dataset

Memory utilization is 8GB RAM for a 500-user analysis

Scalability is Linear performance degradation up to 5,000 users [26]

These results collectively demonstrate the effectiveness of the proposed NLP and behavioral analysis framework for developing accurate, dynamic personas that significantly improve requirements quality in data warehouse environments.[27]

5. RESULT ANALYSIS

This section provides a comprehensive analysis of the experimental results, comparing the proposed NLP and behavioral analysis framework against traditional methods. The analysis is structured across key performance dimensions with statistical validation.

An easy way to comply with the conference paper formatting requirements is to use this document as a camera-ready template.

TABLE III: Comparative Analysis of Requirements Quality Metrics

Quality Dimension	Proposed Method	Traditional Method	Improvement	Statistical Significance	Effect Size (Cohen's d)
Completeness	92.3% (±3.2%)	73.8% (±5.7%)	+25.1%	$p < 0.001$	1.24
Accuracy	88.7% (±4.1%)	71.2% (±6.3%)	+24.6%	$p < 0.001$	1.18
Specificity	94.1% (±2.8%)	76.5% (±5.9%)	+23.0%	$p < 0.001$	1.32
Consistency	89.8% (±3.5%)	74.3% (±6.1%)	+20.9%	$p = 0.002$	0.96
Traceability	91.5% (±3.1%)	69.8% (±7.2%)	+31.1%	$p < 0.001$	1.41

In Table III, the proposed method shows a statistically significant improvement in all quality dimensions, particularly with a substantial effect on detection and perfection. This reflects better alignment between user requirements and final specifications. [29]

TABLE IV: Persona Effectiveness and User Segment Identification

Persona Attribute	Identification Accuracy	Cluster Quality (Silhouette Score)	Evolution Tracking Precision	User Coverage
Executive Consumers	95.2%	0.78	91.8%	98%
Analytical Power Users	89.7%	0.82	88.3%	95%
Operational Reporters	93.4%	0.75	94.1%	97%
Data Explorers	87.9%	0.71	85.6%	92%
Support Users	91.5%	0.69	89.7%	94%
Overall Average	91.5% (±3.1%)	0.75 (±0.05)	89.9% (±3.4%)	95.2% (±2.3%)

Executive Consumers	95.2%	0.78	91.8%	98%

In Table IV, The proposed method shows statistically significant improvement in all quality dimensions, especially on identification and perfection. This reflects better alignment between user's requirements and final specifications. [30]

TABLE V: Efficiency and Performance Metrics Comparison

Performance Metric	Proposed Framework	Baseline Methods	Improvement	Resource Utilization
Requirements Elicitation Time	3.2 days (±0.8)	5.5 days (±1.2)	-42%	35% CPU, 8GB RAM
Revision Cycles	1.3 (±0.4)	3.9 (±1.1)	-67%	-
User Satisfaction (SUS)	85.4 (±4.2)	68.2 (±7.8)	+25.2%	-
Data Processing Throughput	15,000 rec/min	4,500 rec/min	+233%	72% CPU, 12GB RAM
Persona Generation Latency	<30 minutes	>120 minutes	-75%	45% CPU, 8GB RAM

In Table V, the framework demonstrates substantial efficiency gains, particularly in processing throughput and persona generation latency.[10] The reduction in revision cycles indicates more accurate initial requirements capture.[32]

TABLE VI: NLP Component Performance Analysis

NLP Technique	Precision	Recall	F1-Score	Contribution to Persona Accuracy
Topic Modeling (LDA)	0.87	0.83	0.85	28.5%
Sentiment Analysis	0.92	0.88	0.90	19.2%
Named Entity Recognition	0.85	0.79	0.82	15.7%
Syntax Parsing	0.78	0.82	0.80	12.3%
Semantic Similarity	0.91	0.86	0.88	24.3%

In Table V, emotional analysis and economical equality techniques made the most important contribution to personality accuracy, which highlighted the importance of understanding emotional reference and meaning in user needs. [31]

TABLE VII: Statistical Significance of Key Findings

Hypothesis Tested	Test Statistic	p-value	Confidence Interval	Conclusion
H1: Improved Requirements Quality	$t(38) = 6.82$	< 0.001	[15.3%, 28.7%]	Strongly Supported
H2: Enhanced User Satisfaction	$t(38) = 5.91$	< 0.001	[12.8%, 29.6%]	Strongly Supported
H3: Reduced Elicitation Time	$t(38) = -7.23$	< 0.001	[-2.9, -1.7 days]	Strongly Supported
H4: Better Persona Accuracy	$\chi^2(4) = 18.76$			
Hypothesis Tested	Test Statistic	p-value	Confidence Interval	Conclusion

6. CONCLUSION

This research presents and validates a comprehensive framework that integrates Natural Language Processing (NLP) and behavioral analysis for developing data-driven personas in data warehouse requirements engineering. The study demonstrates that traditional approaches to persona development, often reliant on anecdotal evidence and static user representations, are insufficient for capturing the complex, evolving needs of data warehouse users. The experimental results confirm that the proposed framework achieves significant improvements across multiple dimensions.[2] The integration of behavioral data analysis with NLP techniques enabled the identification of five distinct user segments with 91.5% accuracy, each characterized by unique requirements patterns and interaction behaviors. The dynamic persona generation algorithm successfully tracked user evolution over time, with 89.9% precision in predicting requirement changes. Most notably, requirements developed using data-driven personas showed 25.1% higher completeness and 24.6% better accuracy compared to traditional methods, while reducing elicitation time by 42%. The framework's effectiveness was statistically validated through rigorous testing, with all primary hypotheses receiving strong support ($p < 0.001$). The large effect sizes (Cohen's $d > 0.80$) indicate not only statistical significance but substantial practical value for real-world applications.

A. Limitations and Future Research Directions

While this research demonstrates significant advances, several limitations suggest directions for future work:

Domain Specificity: The current framework was validated in a single organizational context. Future research should explore applicability across different industries and data warehouse architectures.

Scalability Constraints: Although the framework scales linearly to 5,000 users, further optimization is needed for very large-scale deployments exceeding 10,000 users.

Real-time Processing: The current implementation uses batch processing. Future work could explore real-time persona generation and requirement adaptation.

Cross-cultural Validity: The study did not examine cultural factors that might influence user behavior and requirement patterns in global organizations.

7. ACKNOWLEDGMENT

This work was supported by the School of Computer Science and Applications, IIMT University, Meerut, India.

REFERENCES

[1] Bazoge, A., Morin, E., Daille, B., & Gourraud, P.-A. (2023). Applying natural language processing to textual data from clinical data warehouses: Systematic review. *JMIR Medical Informatics, 11*, e42477. https://doi.org/10.2196/42477

[2] Droste, J., Deters, H., Puglisi, J., & Klünder, J. (2023, September). Designing end-user personas for explainability requirements using mixed methods research. *IEEE International Requirements Engineering Conference Workshops*, 129–135. https://doi.org/10.1109/rew57809.2023.00028

[3] Eido, W. M., & Ibrahim, I. M. (2025). Analyzing textual data in behavioral science with natural language processing. *Engineering and Technology Journal, 10*(4), 4365–4385. https://doi.org/10.47191/etj/v10i04.02

[4] Feuerriegel, S., Maarouf, A., Bär, D., Geissler, D., Schweisthal, J., Pröllochs, N., Robertson, C. E., Rathje, S., Hartmann, J., Mohammad, S. M., Netzer, O., Siegel, A. A., Plank, B., & Van Bavel, J. J. (2025). Using natural language processing to analyse text data in behavioural science. *Nature Reviews Psychology*. https://doi.org/10.1038/s44159-024-00392-z

[5] Hang Guo and Khasfariyati Binte Razikin. 2015. Anthropological User Research: A Data-Driven Approach to Personas Development. In Proceedings of the Annual Meeting of the Australian Special Interest Group for Computer Human Interaction (OzCHI '15). Association for Computing Machinery, New York, NY, USA, 417–421. https://doi.org/10.1145/2838739.2838816

[6] Huang, L., Jaharadak, A. A., Ahmad, N. I., & Wang, J. (2025). A hybrid decomposition and deep learning model for photovoltaic power forecasting under variable meteorological conditions. *International Journal of Data Warehousing and Mining, 21*(1), 1–22. https://doi.org/10.4018/IJDWM.388673

[7] Jansen, B. J., Salminen, J., Jung, S.-G., & Guan, K. (2021). Creating data-driven personas. In *Data-Driven Personas*. Springer, Cham. https://doi.org/10.1007/978-3-031-02231-9_4

[8] Jansen, B. J., Jung, S.-G., Nielsen, L., Guan, K. W., & Salminen, J. (2022). How to create personas: Three persona creation methodologies with implications for practical employment. *Pacific Asia Journal of the Association for Information Systems, 14*, 1–28. https://doi.org/10.17705/1pais.14301

[9] Khurana, D., Koli, A., Khatter, K., & Singh, S. (2023). Natural language processing: State of the art, current trends and challenges. *Multimedia Tools and Applications, 82*, 3713–3744. https://doi.org/10.1007/s11042-022-13428-4

[10] Lee YH, Choi H, Lee SK (2025), Development of Personas and Journey Maps for Artificial Intelligence Agents Supporting the Use of Health Big Data: Human-Centered Design Approach JMIR Form Res 2025;9:e67272.doi: 10.2196/67272

[11] Li, Y., Meng, C., Tian, J., Fang, Z., & Cao, H. (2024). Data-driven customer online shopping behavior analysis and personalized marketing strategy. *Journal of Organizational and End Use Computing, 36*(1), 1–22. https://doi.org/10.4018/JOEUC.346230

[12] Liang, H., Muhammad, S., & Zainudin, M. S. (2024). Data-driven user personas in requirement engineering with NLP and behavior analysis. *JOIV: International Journal on Informatics Visualization, 8*(4), 2033. https://doi.org/10.62527/joiv.8.4.3625

[13] Märtin, C., Bissinger, B. C., & Asta, P. (2021). Optimizing the digital customer journey—Improving user experience by exploiting emotions, personas and situations for individualized user interface adaptations. *Journal of Consumer Behaviour, 22*(5), 1050–1061. https://doi.org/10.1002/cb.1964

[14] McDonough, S., Adamovic, D., & Kock, N. (2021). Emotion in technostress: An affective events model. *Journal of Organizational and End User Computing, 33*(1), 1–21. https://doi.org/10.4018/joeuc.20210101

[15] McGinn, J. J., & Kotamraju, N. (2008). Data-driven persona development. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1521–1524). Association for Computing Machinery. https://doi.org/10.1145/1357054.1357292

[16] Necula, S.-C., Dumitriu, F., & Greavu-Șerban, V. (2024). A systematic literature review on using natural language processing in software requirements engineering. *Electronics, 13*(11), 2055. https://doi.org/10.3390/electronics13112055

[17] Perwej, Y., Bhuvaneswari, E., Kumar, S., Arulkumar, V., & Nancy, P. (2022, March). Unsupervised feature learning for text pattern analysis with emotional data collection: A novel system for big data analytics. *IEEE International Conference on Advanced Computing Technologies and Applications (ICACTA)*, 1–6. https://doi.org/10.1109/icacta54488.2022.9753501

[18] Rehse, J. R., Abb, L., Berg, G., Feja, S., Hagedorn, P., & Schäfer, T. (2024). User behavior mining. *Business & Information Systems Engineering, 66*, 799–816. https://doi.org/10.1007/s12599-023-00848-1

[19] Rudniy, A. (2022). Data warehouse design for big data in academia. *Computers, Materials & Continua, 71*(1), 979–992. https://doi.org/10.32604/cmc.2022.016676

[20] Sabetzadeh, M., & Arora, C. (2024, January 3). *Practical guidelines for the selection and evaluation of

natural language processing techniques in requirements engineering*. arXiv. https://arxiv.org/abs/2401.01508

[21] Sabetzadeh, M., & Arora, C. (2025). Practical guidelines for the selection and evaluation of natural language processing techniques in requirements engineering. In A. Ferrari & G. Ginde (Eds.), *Handbook on natural language processing for requirements engineering*. Springer, Cham. https://doi.org/10.1007/978-3-031-73143-3_15

[22] Salim, F. A., Ali, H., & Falk, K. (2021). *Towards a data processing framework for enhanced user centric system engineering*. USN Open Archive. https://openarchive.usn.no/usn-xmlui/handle/11250/2987949

[23] Salminen, J., Jung, S.-G., & Jansen, B. J. (2022). Are data-driven personas considered harmful? Diversifying user understandings with more than algorithms. *Persona Studies, 7*(1), 48–63. https://doi.org/10.3316/informit.352977339951659

[24] Salminen, J., Jung, S.-G., Nielsen, L., Şengün, S., & Jansen, B. J. (2022). How does varying the number of personas affect user perceptions and behavior? Challenging the 'small personas' hypothesis! *International Journal of Human-Computer Studies, 168*, 102915. https://doi.org/10.1016/j.ijhcs.2022.102915

[25] Setlur, V. (2025). Supporting human-centric data exploration through semantics and natural language interaction. In *Companion of the 2025 International Conference on Management of Data* (pp. 851–854). Association for Computing Machinery. https://doi.org/10.1145/3722212.3725628

[26] Soon-Gyo Jung, Joni Salminen, Kholoud Khalil Aldous, Bernard J. Jansen, PersonaCraft: Leveraging language models for data-driven persona development, International Journal of Human-Computer Studies, Volume 197, 2025, 103445, ISSN 1071-5819, https://doi.org/10.1016/j.ijhcs.2025.103445.

[27] Spadacini, D. (2024). Navigating change and driving innovation: Leveraging big data for enhanced user behavior analysis and strategic decision-making. *International Journal of Data Science and Big Data Analytics, 4*(2), 16–48. https://doi.org/10.51483/ijdsbda.4.2.2024.16-48

[28] Spielhofer, C. (2025). *Potentials of topic modeling and sentiment analysis for data-driven persona generation* [Master's thesis, University of Applied Sciences Campus 02]. opus.campus02.at. https://doi.org/10.58023/1156

[29] Tan, K.L.; Lee, C.P.; Lim, K.M. A Survey of Sentiment Analysis: Approaches, Datasets, and Future Research. Appl. Sci. 2023, 13, 4550. https://doi.org/10.3390/ app13074550

[30] Tiersen, F., Batey, P., Harrison, M. J., Naar, L., Serban, A. I., Daniels, S. J., Calvo, R. A., & Crawford, T. J. (2021). Smart home sensing and monitoring in households with dementia: User-centered design approach. *JMIR Aging, 4*(3), e27047. https://doi.org/10.2196/27047

[31] Yoon Heui Lee, Hanna Choi, Soo-Kyoung Lee, Development of Personas and Journey Maps for Artificial Intelligence Agents Supporting the Use of Health Big Data: Human-Centered Design Approach, JMIR Formative Research,Volume 9, 2025, ISSN 2561-326X, https://doi.org/10.2196/67272.

[32] Zhao, L., Alhoshan, W., Ferrari, A., Letsholo, K. J., Ajagbe, M. A., Chioasca, E., & Batista-Navarro, R. T. (2021). Natural language processing for requirements engineering. *ACM Computing Surveys, 54*(3), 1–41. https://doi.org/10.1145/3444689