Enhancing Data Warehouse Queries Through Intelligent Keyword Search With Privacy & Access Control

Baalasubramani V[1], Shivanad N[2], Rohith B[3], Subramani V[4]
[1, 2, 3, 4] Dept of Computer science and Engineering
[1, 2, 3, 4] Mahendra Institute of Engineering and Technology, Namakkal, Tamil Nadu, India

Abstract- *Traditional data warehouse systems require users to write structured SQL queries to retrieve information, which is difficult for non-technical users. Moreover, sensitive organizational data stored in warehouses demands controlled access and privacy preservation. Existing keyword search approaches provide basic retrieval but fail to enforce secure role-based data visibility.This paper proposes a Privacy-Preserving Intelligent Keyword Search system integrated with Role-Based Access Control (RBAC) for cloud-based data warehouses. The system converts natural language keywords into structured database queries and dynamically filters results based on user authorization levels. Public data is accessible to all users while confidential records remain restricted to administrators. The proposed model improves usability, security, and controlled data sharing in enterprise environments.*

Keywords: Data Warehouse, Keyword Search, RBAC, Privacy Preservation, SQL Generation, Access Control

I. INTRODUCTION

Modern organizations generate massive volumes of structured data which are stored in centralized data warehouses for analysis and decision-making. Accessing this data typically requires Structured Query Language (SQL), which demands technical knowledge. Non-technical users such as managers, analysts, and administrative staff often struggle to retrieve required information due to the complexity of query syntax.

To overcome this limitation, keyword-based search mechanisms have been introduced where users can retrieve information using simple natural language phrases. However, traditional keyword search systems in data warehouses mainly focus on usability and retrieval accuracy, while ignoring security and privacy. Sensitive organizational data such as financial reports, employee records, and internal analytics should not be visible to all users. Unauthorized access may lead to information leakage and serious security threats.

Cloud-based data warehouses further increase this risk because data is shared across multiple departments and users. Therefore, there is a strong need for a system that not only simplifies data retrieval but also enforces strict access control policies.

In this paper, we propose a Privacy-Preserving Intelligent Keyword Search system with Role-Based Access Control (RBAC). The proposed system converts user-entered natural language keywords into structured queries and filters results based on user roles such as administrator and normal user. Public records are accessible to all users, whereas confidential data is restricted to authorized roles only.

The proposed approach improves usability, enhances data privacy, and ensures secure information retrieval in cloud data warehouse environments.

II. LITERATURE SURVEY

Several research works have focused on improving information retrieval in data warehouse environments. Early systems relied completely on structured SQL queries, which required users to understand database schema and syntax. This created usability issues for non-technical users.

Keyword-based database search techniques were later introduced to allow users to retrieve information using simple text queries. Systems such as relational keyword search and semantic query mapping convert user keywords into SQL statements automatically. These approaches improved usability but lacked access control mechanisms, allowing all users to view all data regardless of sensitivity.

Some studies proposed secure searchable encryption and privacy-preserving search over outsourced databases. These methods protect data from external attackers but do not differentiate between authorized internal users. As a result, confidential data may still be visible to employees without proper permission.

Role-Based Access Control (RBAC) models have been widely used in enterprise applications to restrict system access based on user roles. However, RBAC is generally applied after query execution and not integrated with keyword-to-query translation. Therefore, sensitive data may still be partially exposed during search operations.

From the literature, it is clear that existing systems either focus on usability or security individually, but not both together. There is a need for a unified system that performs intelligent keyword search while simultaneously enforcing privacy and role-based data visibility.

III. PROBLEM STATEMENT

Traditional data warehouse systems primarily depend on Structured Query Language (SQL) for data retrieval, which requires technical expertise and familiarity with database schemas. This creates a significant barrier for non-technical users such as managers and business analysts, leading to inefficient and limited access to critical organizational data.

To address usability issues, keyword-based search techniques have been introduced, allowing users to retrieve data using simple natural language inputs. However, most of these systems lack proper access control mechanisms, resulting in all matching records being displayed regardless of their sensitivity level. This creates a major security concern, as confidential organizational data such as financial reports and internal analytics may become accessible to unauthorized users.

On the other hand, existing secure data retrieval systems focus heavily on access control and data protection but often require complex query structures, reducing usability for general users. Thus, current approaches either prioritize ease of access or data security, but fail to effectively integrate both aspects within a single system.

Therefore, there is a critical need for a unified framework that enables intuitive keyword-based data retrieval while enforcing strict access control policies.

The proposed solution addresses this gap by automatically translating user-entered keywords into structured database queries and dynamically filtering results based on user authorization levels. This ensures that sensitive information remains protected while publicly accessible data can be retrieved efficiently.

IV. PROPOSED SYSTEM

The proposed system presents a Privacy-Preserving Intelligent Keyword Search framework integrated with Role-Based Access Control (RBAC) for secure and efficient data retrieval in cloud data warehouse environments. The primary objective of this system is to eliminate the complexity of SQL-based querying while ensuring that sensitive organizational data remains protected from unauthorized access.

Unlike traditional approaches, the system enables users to retrieve information using natural language keyword queries, thereby improving usability for non-technical users. At the same time, it incorporates a robust access control mechanism to enforce data privacy during the retrieval process.

The overall architecture of the system is divided into four key functional modules: User Authentication Module, Keyword Processing Module, Query Generation Module, and Access Control Module.

Initially, the User Authentication Module verifies the identity of the user and determines the corresponding role, such as administrator or normal user. This role information plays a critical role in controlling access to sensitive data during subsequent operations.

Once authenticated, the user inputs a keyword-based query. The Keyword Processing Module analyzes the input using text parsing techniques to extract meaningful entities such as company name, report type (e.g., salary or profit), and temporal attributes like year. This structured interpretation of unstructured input enables accurate mapping to database fields.

The extracted parameters are then passed to the Query Generation Module, which dynamically constructs the corresponding SQL query. This automated translation eliminates the need for users to manually write SQL statements, thereby significantly enhancing system usability.

After executing the generated query on the warehouse database,the retrieved results are forwarded to the Access Control Module, where RBAC policies are enforced. The system categorizes data into public and private access levels. Public records are accessible to all users, while confidential records are restricted to administrators only. If a normal user attempts to access restricted data, the system prevents access and returns an appropriate “Access Denied” response.

This integrated design ensures that the system achieves a balance between usability and security. By combining intelligent keyword-based query processing with role-based data filtering, the proposed system provides a secure, user-friendly, and efficient solution for controlled information retrieval in cloud data warehouse environments.

A. Module Description

The proposed system is divided into four major functional modules to ensure efficient and secure data retrieval from the cloud data warehouse.

Module 1- User Authentication Module

This module is responsible for validating user credentials. It verifies the username and password entered by the user and determines the user role (Administrator or Normal User). This role identification is essential for enforcing access control policies.

Module 2- Keyword Processing Module

This module processes the user’s natural language query. It applies basic text processing techniques such as tokenization, stop word removal, and keyword extraction to identify important entities like company name, report type, and year.

Module 3- Query Generation Module

This module converts the extracted keywords into a structured SQL query dynamically. It maps the identified entities to database fields and generates appropriate SQL statements to retrieve relevant data from the warehouse.

Module 4- Access Control Module

This module enforces Role-Based Access Control (RBAC). It verifies whether the user has permission to access the requested data. Public records are accessible to all users, while private records are restricted to administrators. Unauthorized access attempts are blocked, ensuring data security.

V. SYSTEM ARCHITECTURE

The proposed system follows a layered and modular architecture designed to ensure both efficient data retrieval and secure access control within a cloud data warehouse environment. The overall workflow is organized as a sequential pipeline, where each module performs a dedicated function in processing user queries.

The architecture begins with the User Layer, where users such as administrators and normal users interact with the system through a web-based interface. The Web Interface, implemented using Flask, serves as the communication bridge between the user and backend processing modules.

Once a request is initiated, the system passes through the Authentication Module, which validates user credentials and determines the user role. This role information is essential for enforcing access control policies in later stages.

The input query is then forwarded to the Keyword Processing Engine, which analyzes the natural language input using text parsing techniques. It extracts key entities such as company name, report type, and temporal attributes, transforming unstructured input into a structured format.

The processed data is then handled by the SQLQuery Generator, which dynamically constructs the appropriate SQL query based on the extracted parameters. This automation eliminates the need for manual query writing.

Before retrieving data from the database, the system applies the RBAC Access Controller, which acts as a security enforcement layer. This module verifies whether the user has sufficient permissions to access the requested data. It ensures that public data is accessible to all users while restricting confidential data to authorized roles only.

The query is then executed on the Cloud Data Warehouse Database, which stores structured organizational datasets. Finally, the Result Display Module presents filtered results to the user. If the user lacks sufficient permissions, an appropriate “Access Denied” message is returned instead of the data.

This structured flow ensures that the system maintains a balance between usability and security by integrating intelligent query processing with strict access control enforcement.

V. ALGORITHMS USED

The proposed system integrates multiple algorithms and techniques to ensure efficient, user-friendly, and secure data retrieval from the data warehouse. Each algorithm plays a specific role in transforming user input into meaningful and authorized results.

1. Keyword Processing Algorithm

The system utilizes a Natural Language Processing (NLP)-based keyword processing algorithm to analyze the user's input query. Initially, the input text is converted into lowercase to maintain uniformity. Stop words such as "the", "of", and "and" are removed to improve relevance. The remaining text is tokenized into individual keywords.

Example:

Input Query: "TCS company profit report 2025"
Processed Output: [tcs, profit, 2025]
This step ensures that only meaningful keywords are considered for further processing.

2. Rule-Based Entity Extraction Algorithm

After preprocessing, a rule-based entity extraction algorithm identifies important entities from the keywords. The system detects elements such as company name, report type, and year using predefined matching rules.

Example:

Keywords: [tcs, profit, 2025]
Company Name → TCS
Report Type → Profit
Year → 2025
These extracted entities are used to generate a structured database query.

3. Dynamic SQL Query Generation Algorithm

The system automatically generates SQL queries based on the extracted entities. This dynamic query generation eliminates the need for manual SQL writing and ensures accurate data retrieval.

Example:

Generated SQL Query:
SELECT * FROM warehouse
WHERE company = 'tcs' AND report_type = 'profit' AND year = 2025;
This allows the system to retrieve relevant data directly from the database.

4. Role-Based Access Control (RBAC) Algorithm

To ensure data security, the system applies a Role-Based Access Control (RBAC) algorithm. Each user is assigned a role such as administrator or normal user. The system checks access permissions before displaying the data.

Example:

Query: "TCS profit report 2025"
If record access = private
Admin → Access granted
Normal User → Access denied
This mechanism prevents unauthorized users from accessing confidential data.

5. Data Filtering Algorithm

After retrieving data from the database, a filtering algorithm is applied to ensure that only authorized and relevant records are displayed. Unauthorized records are removed based on user role and access type.

Example:

Database contains both public and private records
Normal user → Only public data displayed
Admin → Both public and private data displayed
This ensures that the final output is both secure and accurate.

VI. IMPLEMENTATION AND SYSTEM DESIGN

The proposed system is implemented as a web-based application using Python and lightweight web technologies. The overall implementation is structured into three major components: front-end interface, back-end processing, and database management. These components work together to provide efficient and secure data retrieval.

1. Front-End Interface

The user interface is developed using HTML and CSS to provide a simple and interactive environment for users. It allows users to log in and submit keyword-based queries using natural language. The system displays the retrieved results in a structured tabular format. In cases where access is restricted, an appropriate "Access Denied" message is shown to the user.

2. Back-End Processing

The back-end logic is implemented using the Flask framework in Python. Flask handles core functionalities such as user authentication, session management, keyword processing, SQL query generation, and role verification. The system dynamically processes user input and ensures secure communication between the interface and the database.

3. Database Management

The data warehouse is implemented using an SQLite database. It stores structured organizational data, including company name, report type, year, data description, and access level (public or private). The database is accessed through dynamically generated SQL queries, enabling efficient and accurate data retrieval.

4. Keyword Processing Module

The system incorporates basic Natural Language Processing (NLP) techniques such as tokenization and pattern matching to process user queries. Meaningful entities are extracted from the input and mapped to corresponding database fields. This enables automatic conversion of natural language queries into structured SQL statements.

5. Access Control Mechanism

A Role-Based Access Control (RBAC) mechanism is implemented using session-based authorization. The system verifies the user role before granting access to data. Public records are accessible to all users, whereas private records are restricted to administrators. Unauthorized access attempts are handled by displaying an appropriate restriction message.

VII. RESULTS AND DISCUSSION

1. Secure authentication interface of the Cloud Data Warehouse system enabling role-based user login.

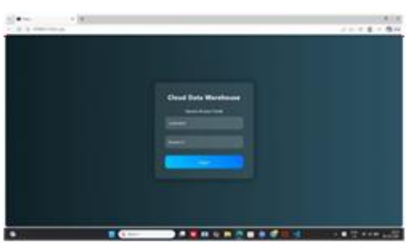

Fig. 1 Login Interface of the Proposed System

The login interface serves as the entry point of the proposed Cloud Data Warehouse system. It provides a secure authentication mechanism that requires users to enter valid credentials, including a username and password, to access the system. Upon successful verification, the system identifies the user role, such as Administrator or Normal User, and grants access based on predefined Role-Based Access Control (RBAC) policies. Unauthorized users are restricted from accessing sensitive organizational data, thereby ensuring data confidentiality and overall system security.

2.Keyword search module allowing users to retrieve warehouse data based on query input and user role.

Fig. 2 – Keyword-Based Data Retrieval Interface

After successful authentication, users are redirected to the keyword search interface. This module enables users to retrieve data from the cloud data warehouse using natural keyword queries such as company name, report type, and year. The system processes the entered keywords and converts them into structured SQL queries to fetch matching records from the database. The access permissions are determined based on the logged-in user role. Admin users can search and retrieve both public and private datasets, whereas normal users are restricted to public data only.

3. Keyword-Based Data Retrieval with Access Control

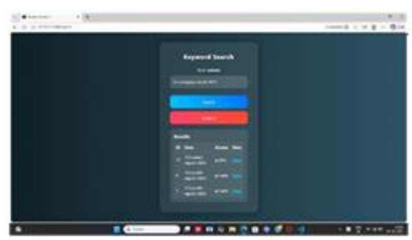

Fig. 3 Keyword-Based Data Retrieval Interface

The figure illustrates the keyword-based search interface of the proposed system after successful user authentication. The user enters a natural language query such as "TCS company report 2025" to retrieve relevant data from the cloud warehouse.

The system processes the query using keyword analysis techniques and converts it into a structured SQL query. The retrieved results are displayed in a tabular format containing Record ID, data description, access level, and view option.

Role-Based Access Control (RBAC) is applied to ensure secure data access. Administrators can access both public and private records, while normal users are restricted to public data only.

This demonstrates efficient and secure keyword-based data retrieval with controlled access to sensitive information.

4.Authorized Record View

Fig. 4 Record Detail View (Authorized Access)

The figure illustrates the detailed record view displayed after selecting a result from the keyword search interface. When the user clicks the "View" option, the system performs role verification and access control validation before retrieving the data.

Since the logged-in user has administrator privileges, the system grants access to private records and displays complete information from the data warehouse. The record details include Record ID, company name, report type, year, access level, and full data content.

The backend enforces authorization checks and executes a secure SQL query to fetch the requested record. This ensures that sensitive organizational data is accessed only by authorized users.

This figure demonstrates the secure data retrieval mechanism of the proposed system, where confidential data is protected while being accessible to privileged roles.

5.Unauthorized Access Control

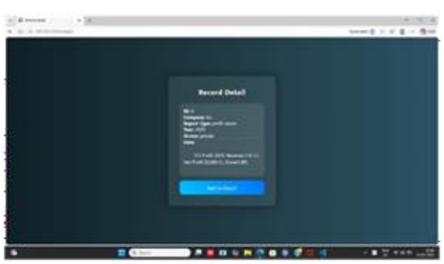

Fig. 5 Unauthorized Access Blocked (User Role)

The figure illustrates the enforcement of Role-Based Access Control (RBAC) in the proposed system when an unauthorized user attempts to access a private record.

When a normal user tries to view a restricted dataset, the system verifies access permissions before retrieving data. Since the requested record is marked as private and the user lacks sufficient privileges, access is denied and an appropriate message is displayed.

The system prevents exposure of sensitive information and redirects the user back to the search interface without transmitting any confidential data.

This demonstrates that the system effectively differentiates user roles, restricts unauthorized access, and ensures data security by preventing information leakage in the cloud data warehouse environment.

VIII. PERFORMANCE ANALYSIS AND ADVANTAGES

The proposed Role-Based Secure Cloud Data Warehouse system was evaluated using multiple user roles and datasets consisting of both public and private enterprise records. The system successfully performed user authentication, keyword-based query processing, and controlled data retrieval based on access permissions.

During testing, the administrator was able to access both public and private organizational data, including confidential reports such as company profit records. In contrast, when a normal user attempted to access private datasets, the system denied access and prevented exposure of sensitive information. However, public datasets such as salary reports remained accessible to all authorized users.

The keyword-based search mechanism efficiently converts user queries into structured SQL conditions and retrieves only relevant and authorized data. This improves both system efficiency and data security by avoiding unnecessary data processing.

A. Advantages of the Proposed System

1. Secure Data Access

The RBAC mechanism ensures that sensitive organizational data is accessible only to authorized users, preventing data leakage.

2. Controlled Information Sharing

The system separates public and private datasets, enabling safe and selective data sharing within organizations.

3. Efficient Query Processing

Keyword-based queries are dynamically converted into optimized SQL queries, reducing data retrieval time.

4. Unauthorized Access Prevention

The system enforces access control at the server level, preventing unauthorized users from accessing restricted records.

5. Cloud-Ready Architecture

The system is designed to be easily deployable in real cloud environments without significant modification.

6. Improved Data Privacy

Sensitive business data such as financial reports and internal analytics are protected from unauthorized exposure.

IX. SOFTWARE AND HARDWARE REQUIREMENTS

Software Requirements

1. Programming Language: Python

Used for backend development and implementing logic such as authentication, keyword processing, and access control.

2. Framework: Flask

Used to build the web application and handle routing, user sessions, and request processing.

3. Front-End Technologies: HTML, CSS

Used to design user interface including login page, search page, and result display.

4. Database: SQLite

Used as a lightweight database to store company data, reports, and access control information.

5. Development Tool: Visual Studio Code

Used for writing, editing, and running the project code.

6. Browser: Google Chrome / Edge

Used to run and test the web application interface.

Hardware Requirements

1. Processor: Dual Core (Intel i3 or above)
   Required to execute the application smoothly.
2. RAM: Minimum 4 GB
   Required to run Python, Flask server, and browser without lag.
3. Storage: Minimum 256 GB
   Used to store project files, database, and software tools.
4. Internet Connection
   Required if the system is deployed on cloud or for downloading libraries and tools.
5. Display Device: Monitor or Laptop Screen
   Used for interacting with the web interface.
6. Input Devices: Keyboard and Mouse
   Required for user input and system interaction.

X. CONCLUSION

This paper presents a secure cloud data warehouse system integrated with Role-Based Access Control (RBAC) to protect sensitive organizational data in cloud environments. The system combines user authentication, intelligent keyword-based search, and permission-aware data retrieval into a unified framework that enhances both usability and security. It effectively differentiates between administrator and normal user roles, where administrators can access both public and private datasets while normal users are restricted to publicly available information. Unauthorized access attempts are blocked by the system, ensuring confidentiality of critical business data. The keyword-based search mechanism allows users to retrieve information using natural language instead of complex SQL queries, and the system automatically converts these queries into structured database operations. Experimental results demonstrate that the proposed system successfully prevents data leakage, enforces access control policies, and maintains secure data availability for authorized users. Hence, the system provides a practical, efficient, and scalable solution for secure data management in cloud-based data warehouse environments.

REFERENCES

[1] S. Chaudhuri and U. Dayal, “An overview of data warehousing and OLAP technology,” ACM SIGMOD Record, vol. 26, no. 1, pp. 65–74, 1997.
[2] A. Silberschatz, H. F. Korth, and S. Sudarshan, Database System Concepts, 6th ed. New York, NY, USA: McGraw-Hill, 2011.
[3] X. Song, Y. Shi, and S. Yu, “Privacy-preserving keyword search over encrypted data in cloud computing,” IEEE Trans. Cloud Computing, vol. 8, no. 2, pp. 456–469, 2020.
[4] Q. Wang, C. Wang, K. Ren, W. Lou, and J. Li, “Enabling secure and efficient ranked keyword search over outsourced cloud data,” IEEE Trans. Parallel Distrib. Syst., vol. 23, no. 8, pp. 1467–1479, 2012.
[5] J. Li, Q. Wang, C. Wang, N. Cao, K. Ren, and W. Lou, “Fuzzy keyword search over encrypted data in cloud computing,” in Proc. IEEE INFOCOM, 2010, pp. 1–5.
[6] R. Sandhu, E. Coyne, H. Feinstein, and C. Youman, “Role-Based Access Control Models,” IEEE Computer, vol. 29, no. 2, pp. 38–47, 1996.
[7] Flask Documentation, “Flask Web Framework,” Available: https://flask.palletsprojects.com/
[8] SQLite Documentation, “SQLite Database Engine,” Available: https://www.sqlite.org/