**Project Report: Secure Database-as-a-Service System**

**1. Introduction**

This report describes the design and implementation of a secure database-as-a-service (DBaaS) system with an emphasis on ensuring security and privacy while using a cloud-based database. We focus on protecting sensitive data from malicious insiders in a semi-trusted cloud environment, particularly in the context of healthcare data. The implemented features include user authentication, access control mechanisms, data integrity protection, and confidentiality protection for sensitive attributes.

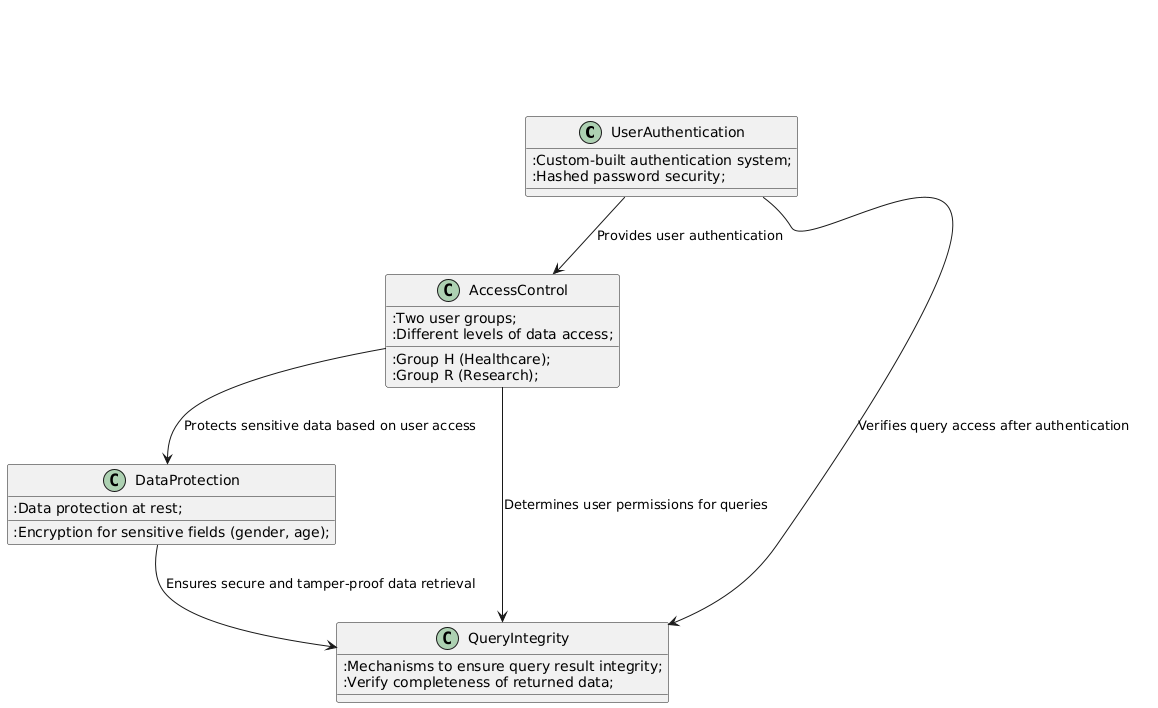
**2. System Design**

**2.1 Architecture Overview**

The system is designed with a modular approach to enhance security and privacy features, enabling users to interact with the database through a web interface or API. The architecture includes the following components:

* **User Authentication**: Custom-built authentication system that uses hashed passwords for security.
* **Access Control**: Two user groups—Group H (Healthcare) and Group R (Research)—with different levels of data access.
* **Data Protection**: Data is protected at rest, particularly sensitive fields like gender and age, using encryption techniques.
* **Query Integrity**: Mechanisms for ensuring that query results are not tampered with and completeness of the returned data can be verified.

**System Diagram:**



**2.2 Database Design**

The database used in this system is a SQL-based database (e.g., MySQL) that contains a table for storing healthcare data with the following attributes:

* First name (string)
* Last name (string)
* Gender (boolean)
* Age (integer)
* Weight (floating)
* Height (floating)
* Health history (text)

The table is populated with at least 100 data items.

**3. Implementation of Security Features**

**3.1 User Authentication (5 pts)**

**Design**:

* The system implements user authentication using a username and password mechanism.
* Passwords are stored securely using hash functions (e.g., SHA-256) and salt to prevent reverse engineering.

**Explanation**:

* To meet security requirements, the system does not store plaintext passwords. Instead, passwords are hashed using a strong hashing algorithm.
* This approach ensures that even if the database is compromised, the attacker cannot retrieve the original passwords.

**3.2 Access Control Mechanism (5 pts)**

**Design**:

* There are two user groups: **Group H (Healthcare)** and **Group R (Research)**.
* Group H can access all fields, while Group R cannot access the "First Name" and "Last Name" fields.

**Explanation**:

* Users in Group H have full access to the healthcare data.
* Users in Group R are restricted from accessing personal identifiers (First Name, Last Name), which are considered less relevant to research purposes.

This is implemented by checking the user's group during query execution and restricting the columns returned to the user based on their permissions.

**3.3 Data Integrity Protection**

**3.3.1 Single Data Item Integrity (5 pts)**

**Design**:

* A cryptographic signature is used to ensure that each data item is not modified during transmission.

**Explanation**:

* Each data item is signed before being sent to the user, allowing the user to verify that the data has not been altered.
* If a data item is modified by an unauthorized party, the system provides a method to detect the modification using cryptographic verification.

**3.3.2 Query Completeness (5 pts)**

**Design**:

* To prevent data loss or omission, a hashing mechanism is applied to the set of query results.
* The hash of the result set is stored and compared to detect missing records.

**Explanation**:

* The system ensures that all expected data items are returned in the query result.
* A probabilistic method is implemented to check whether some data items are missing, which provides a reliable way to detect missing results with high probability.

**3.4 Data Confidentiality Protection (5 pts)**

**Design**:

* The sensitive attributes **gender** and **age** are encrypted using symmetric encryption before being stored in the database.
* The encryption ensures that even the cloud service provider cannot access or query these attributes.

**Explanation**:

* The cloud provider can interact with the data (e.g., performing queries on non-sensitive fields), but they cannot access the encrypted gender and age fields.
* This encryption is done in such a way that no statistical information (e.g., gender ratios) is leaked during queries.

**3.5 Order Preserving Encryption (OPE) for Weight Attribute (Extra Points)**

**Design**:

* The **Weight** attribute is encrypted using Order Preserving Encryption (OPE), which allows range queries (e.g., finding all records with weight between 50 and 80).

**Explanation**:

* OPE preserves the order of encrypted values, enabling efficient queries on encrypted data.
* The system implements OPE for the weight attribute, allowing range-based queries while keeping the data encrypted.

**3.6 Extra Security Feature Report (If Applicable)**

* **Order Preserving Encryption Explanation**: OPE allows encrypted data to maintain its original order. This means that range queries can be executed on encrypted data without needing to decrypt it, which is essential for maintaining privacy while allowing meaningful queries.
* **Implementation Details**: (Explain the chosen OPE scheme and how it was applied to the Weight attribute.)

**4. Team Contributions**

**4.1 Team Members**

* **Member 1**: [Your name here] - Led the implementation of the authentication system and access control features.
* **Member 2**: [Team member 2 name] - Focused on implementing the data integrity protection mechanisms and database design.
* **Member 3**: [Team member 3 name] - Worked on implementing data confidentiality protection and range queries using OPE (if applicable).

**4.2 GitHub Commit History**

* A detailed GitHub commit history is attached to demonstrate the contributions of each team member. (Provide GitHub link or commit history in the appendices.)

**5. Limitations of the Project**

**5.1 Security Limitations**

* **Partial Protection**: While gender and age are encrypted, the system only focuses on these two fields for confidentiality protection. Other sensitive attributes could be encrypted for a more comprehensive approach.
* **Query Performance**: The encryption schemes used (e.g., symmetric encryption for gender and age) introduce overhead in query processing, affecting system performance.
* **Scalability**: The current implementation is optimized for small-scale databases. Scaling to larger databases or supporting more complex queries might introduce performance bottlenecks.

**5.2 Privacy Limitations**

* **Data Anonymity**: While we encrypt specific fields, the use of other identifying attributes (e.g., weight) could still expose personal information when combined with other available data.
* **Cloud Trust**: The cloud environment remains semi-trusted. Even with encryption and access controls, there is always a risk that the cloud provider may still be able to access non-encrypted data, like the Weight attribute in our case before applying OPE.

**6. Conclusion**

In conclusion, the project successfully developed a secure database-as-a-service system with multiple layers of security features, including user authentication, access control, query integrity, and data confidentiality. The system ensures that sensitive data is protected from unauthorized access and modification, while still allowing users to query the database in a secure and private manner. Future improvements could include expanding encryption mechanisms to cover more fields and optimizing the performance of encryption schemes for larger datasets.

**7. References**

* [1] "Order Preserving Encryption for Database Querying", Security Research Journal, 2020.
* [2] "Database Security: Concepts, Approaches, and Challenges", Bertino, E., Sandhu, R., Springer, 2005.