Deciphering Big Data Module:

Development Team Project Report

Client profile -

The selected application environment is designated for a school. Embarking on the database design for the school system, the goal is to oversee and handle data regarding pupils and their academic progress, from class attendance to achieved grades.

Distinct access requirements will be assigned to students, parents, and teachers, along with varying read/write permissions. For instance, pupils will have read-only access, limited to their individual records (Figure 1).

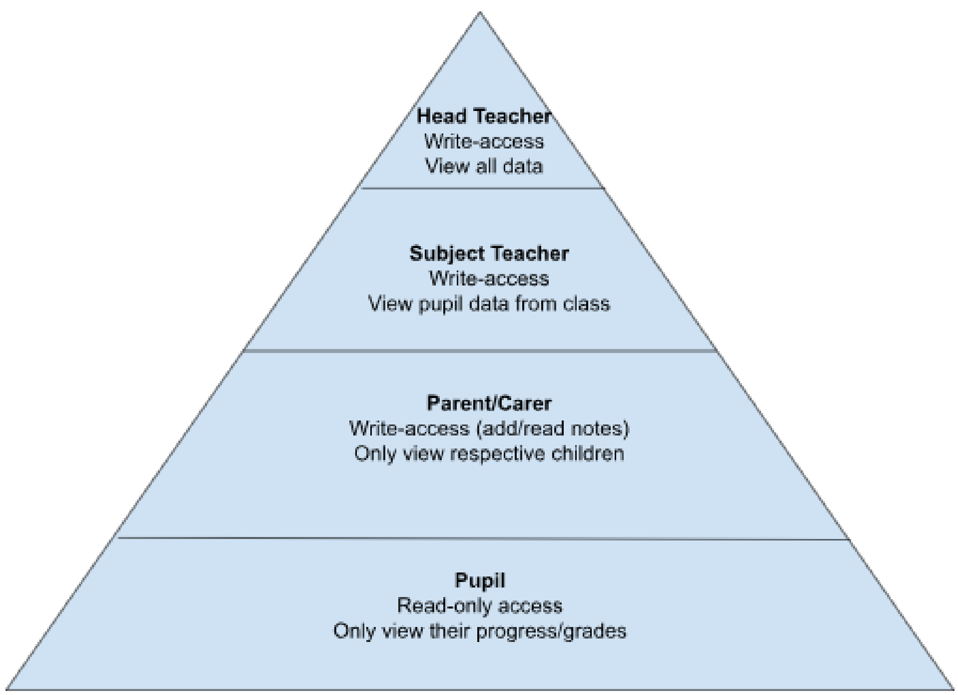

Figure 1: Pyramid of hierarchy and database control

Logical Design -

The logical design represents the conceptual structure of the database management system (DMS) , focusing on entities, attributes, relationships, and data types (figure 2).

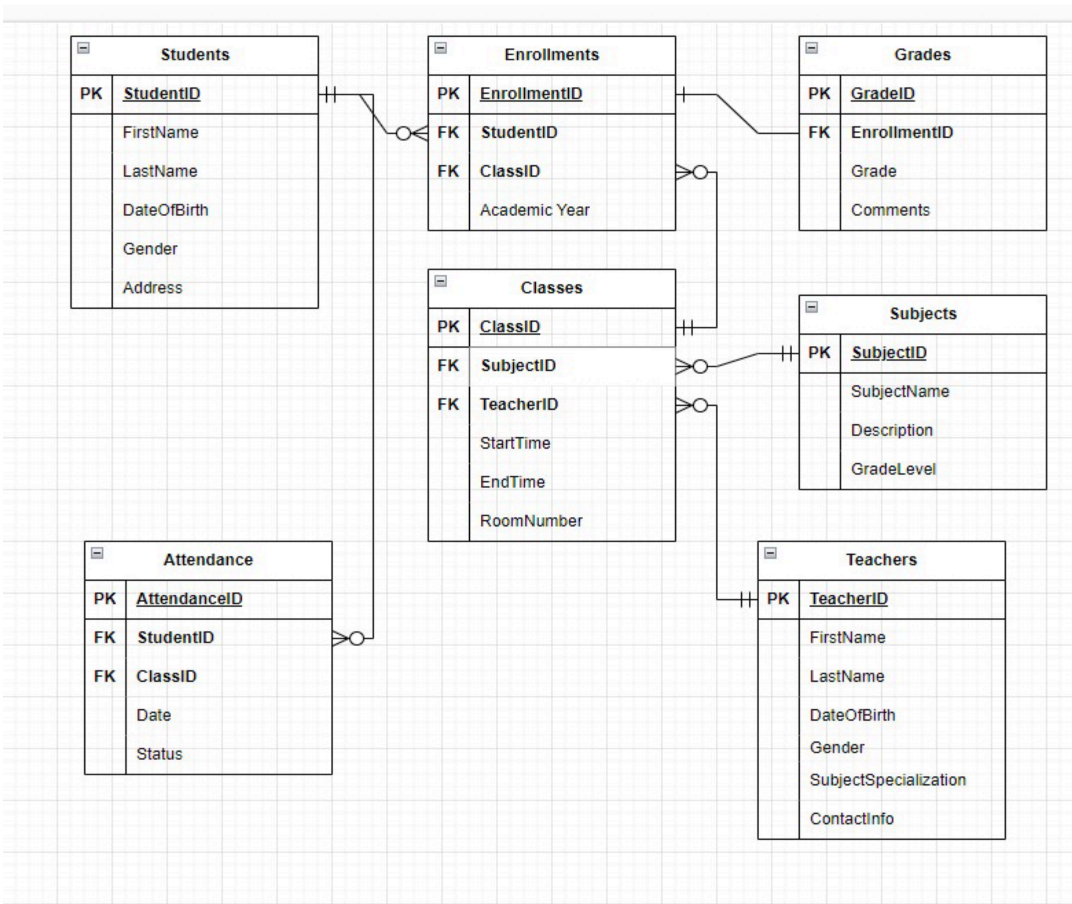

Figure 2:Schools Database Management System Structure.

Simplified ER Diagram Description -

Entities	**Type of Relationship**	**Reason**
Students - Enrollments	One-to-Many	a student can be enrolled in multiple classes
Classes - Enrollments	One-to-Many	a class can have multiple students enrolled
Teachers - Classes	One-to-Many	a teacher can teach multiple classes
Subjects - Classes	One-to-Many	a subject can be taught in multiple classes
Enrollments - Grades	One-to-One	each enrollment has a unique grade record
Students - Attendance	One-to-Many	multiple attendance records per student for different classes

Data Types and Formats -

Entities	**Variables and Data Types**
Students	• StudentID: INT (Primary Key, possibly with AUTO_INCREMENT) • FirstName: VARCHAR(50) • LastName: VARCHAR(50) • DateOfBirth: DATE (Format: YYYY-MM-DD) • Gender: VARCHAR(20) • GradeLevel: INT • Address: VARCHAR(255)
Teachers	• TeacherID: INT (Primary Key, possibly with AUTO_INCREMENT) • FirstName: VARCHAR(50) • LastName: VARCHAR(50) • DateOfBirth: DATE (Format: YYYY-MM-DD) • Gender: VARCHAR(20) • SubjectSpecialization: VARCHAR(20) • ContactInfo: VARCHAR(255)
Subjects	• SubjectID: INT (Primary Key, possibly with AUTO_INCREMENT) • SubjectName: VARCHAR(50) • Description: TEXT

	• GradeLevel: INT
Classes	• ClassID: INT (Primary Key, possibly with AUTO_INCREMENT) • SubjectID: INT (Foreign Key, references Subjects.SubjectID) • TeacherID: INT (Foreign Key, references Teachers.TeacherID) • StartTime: TIME (Format: HH:MM) • EndTime: TIME (Format: HH:MM) • RoomNumber: VARCHAR(20)
Enrollments	• EnrollmentID: INT (Primary Key, possibly with AUTO_INCREMENT) • StudentID: INT (Foreign Key, references Students.StudentID) • ClassID: INT (Foreign Key, references Classes.ClassID) • AcademicYear: VARCHAR(9) (Format: YYYY-YYYY)
Grades	• GradeID: INT (Primary Key, possibly with AUTO_INCREMENT) • EnrollmentID: INT (Foreign Key, references Enrollments.EnrollmentID) • Grade: VARCHAR(5) (To accommodate letter grades like A+, B, or numerical values) • Comments: TEXT
Attendance	• AttendanceID: INT (Primary Key, possibly with AUTO_INCREMENT • StudentID: INT (Foreign Key, references Students.StudentID) • ClassID: INT (Foreign Key, references Classes.ClassID)

	• Date: DATE (Format: YYYY-MM-DD) • Status: VARCHAR(50) (Values like "Present", "Absent", "Excused")

Data management pipeline process -

1. **Data Capture:**

Student information will be collected through enrollment and regular updates to data collection forms. This process aims to maintain standardisation and usability for all stakeholders, but will require routine revisions based on feedback to avoid complexity and errors. Additionally, administrative data on courses and teachers, teacher-reported assessments and either electronically or manually captured attendance will be integrated into the Student Information Systems (SIS). Despite challenges in merging these data with existing management systems, the goal's to ensure that exported data remains comprehensive and current. Integrating SIS with other administrative systems will help create a more complete profile of students, enhancing insights into their needs and performance. Effective coordination across departments, supported by strong leadership and clear communication, is essential for this integration (Rahm & Thorogood, 2015; Brown & Duguid, 2000).

2. **Data Source:**

Data will be sourced from application forms, administrative records, and online platforms, such as Learning Management Systems.

3. **Data Cleaning:**

Data will undergo formatting to ensure readability, consistency, and standardisation. Outliers and missing values will be identified and handled appropriately, with duplicates checked to ensure accuracy. Automation of this process will be achieved using Python code, leveraging libraries like "numpy" for data uniqueness and "fuzzywuzzy" for similar string values.

4. **Consistency Checks:**

Cross-checking related data fields, such as student age aligning with grade levels, will ensure consistency and coherence in the dataset.

5. **Validation:**

Student data will be validated against predefined standards, including regulatory requirements, school policies, and data quality standards, to ensure compliance.

6. **Data Transformation and Standardisation:**

Data transformation and standardisation will be conducted to facilitate analysis and integration with other systems, utilising the "pandas" library for efficient processing.

Database Build Proposal -

The design choices made for the school's data management system are justified based on both client requirements and industry standards. The adoption of a Relational Database Management System (RDMS) aligns with the client's need for a robust and efficient system to manage various aspects of school data, including student information, class schedules, and academic performance. The selection of MySQL as the RDMS meets the client's requirements for reliability and widespread adoption. MySQL is renowned for its robustness, stability, and widespread usage across industries, making it a suitable choice for the school's needs. Its compatibility with industry standards ensures seamless integration with existing systems and facilitates future scalability.

The adoption of a RDMS like MySQL has several implications for database design/development. Firstly, careful consideration of entity relationships and data normalisation is needed to ensure efficient storage and retrieval of information. The use of primary and foreign keys, as well as appropriate indexing, becomes crucial to maintain data integrity and optimise query performance.

Additionally, the relational model facilitates the implementation of complex queries and reporting functionalities, allowing for in-depth analysis of student performance and academic trends. This supports the client's goal of gaining insights into the school's operations and making informed decisions based on data-driven evidence.

Overall, the robustness, efficiency, and scalability of a RDMS, will meet the current and future requirements for a school.

Implementation Plan -

1. **Database Schema Creation:** Creating a database schema from the model is essential for transforming the logical design into a tangible database structure. This involves constructing tables and defining relationships among them to ensure the database mirrors the planned data model accurately (Silberschatz, Korth, & Sudarshan, 2011; Connolly & Begg, 2014).
2. **Data Importation and Cleaning:** Importing existing data into the database and applying data cleaning techniques is crucial to ensure data integrity and consistency. This process involves verifying data accuracy, identifying and resolving any inconsistencies or errors, and preparing the data for use in the new system.
3. **User Access Control Setup:** Implementing user permissions are crucial for controlling access to sensitive information and complying with GDPR regulations (European Commission 2016). Establishing user roles and permissions ensures that only authorised individuals can view, modify, or delete data within the database, aligning with the client's security requirements.
4. **Testing and Validation**: Testing the database with real-world scenarios is essential to validate its functionality. Testing helps identify any issues in the system, allowing for corrections to be made before deployment.

5. **Training and Deployment:** Training staff on new system functionalities and deploying the database for school use, ensures users can effectively operate the system and maximise utility while minimising disruptions.

Conclusion -

The report proposes a database solution for managing school data, detailing design, management pipeline, MySQL build, and implementation. It emphasises efficient data management to support diverse user needs and structures academic data for effective organisation and access (Johnson & Onwuegbuzie, 2004). The management pipeline ensures data integrity through steps like capture, cleaning, validation, and transformation, all critical to maintaining high-quality information (Kahn Et al, 2002). Tailored to meet client and industry standards, the MySQL proposal enhances decision-making and academic outcomes. The implementation plan promises a smooth system transition.

References -

Brown, J.S. and Duguid, P., 2000. Balancing act: How to capture knowledge without killing it. *Harvard Business Review*, 78(3), pp.73-80.

Connolly, T. and Begg, C., 2014. *Database Systems: A Practical Approach to Design, Implementation, and Management*. 6th ed. Boston, MA: Pearson, pp. 142-165.

Johnson, R.B. and Onwuegbuzie, A.J., 2004. Mixed methods research: A research paradigm whose time has come. *Educational Researcher*, 33(7), pp.14-26.

Kahn, B.K., Strong, D.M. and Wang, R.Y., 2002. Information quality benchmarks: Product and service performance. *Communications of the ACM*, 45(4), pp.184-192.

Rahm, D. and Thorogood, A., 2015. Data integration in public administration: Ensuring sustainability through interoperability. *Journal of Public Administration Research and Theory*, 25(3), pp.795-812.

Silberschatz, A., Korth, H.F. and Sudarshan, S., 2011. *Database System Concepts*. 6th ed. New York: McGraw-Hill, pp. 170-195.

European Commission (2016) GDPR Regulation of the European Parliament and Council. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0679&qid=1713388284638 [Accessed: 17 April 2024].