**Assignment 1**

By Harsh Sharma

Part 1

**Part I. Overview of Database Systems**

**Q1** The most classic paper to introduce database systems is M. Stonebraker and J. M. Hellerstein, “*What Goes Around Comes Around*”. Read the paper and answer each of the following questions briefly.

1. What is the notion of *data independence?* Why is it highly desirable?

Solution: The notion of data independence refers to the ability to modify the database schema (the structure or organization of the data) without affecting the applications that use the data.

The reason it’s highly desirable:

* **Flexibility:** It allows for easier adaptation to changing requirements or technological advancements. Developers can modify the database structure or storage mechanisms without disrupting existing applications.
* **Maintainability:** It simplifies maintenance efforts by decoupling the database schema from the applications that use it. This reduces the risk of introducing errors or inconsistencies when making changes to the database

1. What are the key ideas behind the relational model? Why/how it improves previous model?

Solution:

The key ideas behind the relational model were:

* + Store the data in a simple data structure (tables)
  + Access it through a high level set-at-a-time DML
  + No need for a physical storage proposal

The way it improves previous models is by:

 **Flexibility:** The relational model's declarative query language allows users to specify what data they want without needing to know the internal storage structures or access paths. This enhances flexibility and reduces the complexity of application development.

 **Data Independence:** By separating the logical and physical aspects of the database, the relational model provides greater data independence, allowing for easier schema modifications and reducing the impact of changes on applications.

 **Normalization:** The relational model promotes normalization, which improves data integrity, reduces redundancy, and simplifies maintenance compared to the denormalized structures common in hierarchical and network models.

**(3)** What is the motivation of the OR (Object-Relational) model? What are the main additions in ORDB?

Solution: The Object-Relational (OR) model emerged as a response to the limitations of purely relational databases in handling complex data types and relationships. Some of the movtivations were:

* **Handling Complex Data**
* **Supporting Object-Oriented Programming**
* **Enhancing Expressiveness**

The main additions in Object-Relational Database Management Systems (ORDBMS) compared to traditional RDBMS include:

1. **Support for Complex Data Types:** ORDBMSs extend the data types supported by traditional RDBMSs to include complex data types such as arrays, nested tables, and user-defined types. This allows for more flexible data modeling and storage.
2. **Object-Oriented Features:** ORDBMSs incorporate object-oriented features such as classes, objects, and methods, allowing developers to model data more closely to the way it is represented in object-oriented programming languages.
3. **Inheritance and Polymorphism:** ORDBMSs support inheritance relationships between types, allowing attributes and methods to be inherited from parent types. This promotes code reuse and simplifies data modeling.

**(4)** Data can be divided to four categories: rigidly structured data, rigidly structured data with some text fields, semi-structured data, text (unstructured data). Describe each one and give an example for that.

Solution:

* **Rigidly Structured Data:** This category includes data that is highly organized and follows a strict schema or format. Each data element has a predefined data type and is stored in a tabular structure with well-defined rows and columns.

**Example**: A customer database in a relational database management system (RDBMS) where each customer record contains fields such as Customer ID, Name, Address, Phone Number, and Email. Each field has a specific data type (e.g., integer, string, date), and every record follows the same structure.

* **Rigidly Structured Data with Some Text Fields:** Similar to rigidly structured data, this category also follows a strict schema, but it includes text fields that may contain unstructured or semi-structured data. While most of the data adheres to a structured format, some fields allow for free-form text or limited variability.

**Example:** An inventory management system where product descriptions are stored alongside structured data such as Product ID, Name, Price, and Quantity in Stock. The Product Description field allows for varying text lengths and formats, providing additional information about each product.

* **Semi-Structured Data:** Semi-structured data lacks a rigid schema but exhibits some organizational structure. It may contain nested or hierarchical data elements and allows for flexibility in data representation**.**

**Example:** JSON (JavaScript Object Notation) or XML (eXtensible Markup Language) documents used to store configuration settings, web data, or data exchanged between systems. These formats allow for nested structures and key-value pairs, providing flexibility in representing data relationship.

* **Text (Unstructured Data):** This category encompasses free-form textual data that lacks a predefined schema or structure. Text data may include natural language text, documents, emails, articles, or social media posts.

**Example**: A collection of customer reviews for a product stored in a text document or a database field. Each review may contain varying lengths of text, different writing styles, and diverse opinions, making it challenging to extract structured information without natural language processing techniques.

**Q2.** After reading “Section 5” of M. Stonebraker et al. "One Size Fits All", give a short answer on each question following:

1. Does one size fits all? If not, explain the reason.

Solution: No, one size does not fit all. The reason is that different database management systems (DBMS) are designed to address specific use cases and requirements, such as scalability, performance, data model flexibility, or transactional consistency. Trying to apply a single DBMS solution to all scenarios would result in compromises in one or more aspects, leading to suboptimal performance or functionality.

1. What components should be considered in a sensor network system with databases.

Solution: These components may include specialized data storage and processing mechanisms for handling time-series data, efficient data collection and communication protocols for sensor nodes, and robust data management systems capable of handling large volumes of sensor data.

1. Describe the characteristics of (typical) text search system which are different from conventional business-processing applications.

Solution: The typical text search system are:

1. **Unstructured Data Handling:** Text search systems primarily deal with unstructured or semi-structured textual data, unlike conventional business-processing applications that often handle structured data stored in databases.
2. **Scalability for Large Volumes of Text:** These systems must be scalable to handle large volumes of text data efficiently, including indexing and querying vast document collections.
3. **Relevance Ranking:** Unlike conventional applications where queries often have clear-cut answers, text search systems must rank and retrieve relevant documents based on relevance to the user's query.
4. **Full-Text Indexing:** Text search systems rely on full-text indexing techniques to create searchable indexes of document content, enabling fast and efficient retrieval of relevant documents.

**(4)** What should be considered for scientific databases.

Solution: For scientific databases, several factors should be considered to ensure they effectively support the needs of researchers, scientists, and other stakeholders:

1. **Data Types and Formats:** Scientific databases often contain diverse data types and formats, including numerical data, images, text, genomic sequences, molecular structures, and more. The database should support storing and querying various data types effectively.
2. **Metadata Standards:** Standardized metadata is crucial for describing and categorizing scientific data. The database should support metadata standards relevant to the specific scientific domain, facilitating data discovery, sharing, and interoperability.
3. **Data Integrity and Quality:** Scientific data must be accurate, reliable, and of high quality. The database should enforce data integrity constraints, such as validation rules and referential integrity, to ensure data consistency and reliability.

**Q3.** Name the six (/five) phases of System Development Life Cycle (SDLC), and explain the purpose and outputs of each phase.

Solution: The six phases of SDLC are:

**1. Planning Phase:**

**Purpose:** The Planning Phase sets the foundation for the entire project by defining its scope, objectives, timeline, budget, and resources. It involves gathering requirements, assessing feasibility, and creating a project plan.

**Outputs:** The outputs of this phase include a project charter, feasibility study, project plan, requirements documentation, and a risk management plan.

**2. Analysis Phase:**

**Purpose:** The Analysis Phase focuses on understanding the requirements of the system. It involves analyzing existing systems (if any), gathering user requirements, and documenting functional and non-functional requirements.

**Outputs:** The outputs of this phase include a detailed requirements specification document, use cases, system models (e.g., data flow diagrams, entity-relationship diagrams), and a system requirements specification.

**3. Logical Design Phase:**

**Purpose:** The Logical Design Phase translates the requirements gathered in the Analysis Phase into a blueprint for the system. It involves designing the architecture, data model, user interface, and other system components.

**Outputs:** The outputs of this phase include system architecture diagrams, database schemas, user interface designs, system specifications, and prototypes/mockups.

**4.** **Physical Design Phase:**

**Purpose:** The Physical Design Phase involves building the system according to the design specifications. It includes coding, unit testing, integration testing, and deployment activities.

**Outputs:** The primary output of this phase is the implemented system or software, along with documentation such as user manuals, installation guides, and system administration guides.

**5. Implementation Phase:**

**Purpose:** The Implementation Phase verifies that the system meets the specified requirements and functions correctly. It involves various types of testing, including functional testing, performance testing, security testing, and user acceptance testing.

**Outputs:** The outputs of this phase include test plans, test cases, test scripts, test reports, and defect logs.

**6. Maintenance Phase:**

**Purpose:** The Maintenance Phase involves maintaining and supporting the system after it has been deployed. It includes addressing user feedback, fixing defects, implementing enhancements, and ensuring the system's continued reliability and performance.

**Outputs:** The outputs of this phase include maintenance plans, release notes, bug fixes, patches, updates, and documentation updates.