Mid-sem Exam: IIA (CSE656), Monsoon Semester 2025

Time: 1 hr **Marks: 20**

Note: There are Thirty (30) MCQ with a single correct answer. You need to attempt all questions. There is a 50% negative marking if wrong answer is selected. All questions are based on the following case study.

You are NOT allowed to use your Mobile phone. If caught, the case will be reported to the DAC.

SastaBookStore.com is a new start-up that aims to provide a seamless shopping experience by aggregating books from multiple online bookstore platforms into a single, consistent, and coherent interface. Customers should be able to search for books, compare prices, check availability in real time, manage their transaction history, and receive personalized recommendations, all without realizing the data is coming from diverse and heterogeneous sources.

Each bookstore platform exposes its own schema, leading to significant integration challenges. For example, BookStore1 provides a schema with attributes (ISBN, Title, Pub-Year, Availability, Unit_Price, Author, Publisher), while BookStore2 offers (ISBN, BTitle, Year-of-Publication, Inventory, Price, Writer, PublisherName), and BookStore3 exposes (ISBN, Book_Title, Publication-Year, InStock, Price, Authors[], Pub). Beyond these schema differences, each vendor also uses a different data access method: BookStore1 provides a RESTful API, BookStore2 a SOAP-based API, and BookStore3 a flat CSV file dump via FTP. This requires a sophisticated wrapper/adapter layer for the system to abstract these communication protocols, authentication, and rate limiting details. Despite all three storing core book information, attribute names differ (e.g., Title, BTitle, Book_Title), data formats are inconsistent (e.g., prices as floats vs. strings with currency symbols), and some attributes have structural mismatches (BookStore3 stores multiple authors as an array, while others use single-valued attributes). Even publisher names vary (Publisher, PublisherName, Pub), with frequent spelling and formatting inconsistencies. This necessitates a dedicated Master Data Management (MDM) module to perform entity resolution, identifying and merging duplicate entities like "J.K. Rowling" and "Joanne Rowling," and standardizing names using fuzzy matching and phonetic algorithms. The system must track data provenance, linking unified entities back to their source representations.

To address these challenges, the you as a design team member proposes a global unified schema called BookGlobal(ISBN, Title, Authors[], Publisher, PubYear, SubjectArea, AvailabilityStatus, Prices[]). Here, attributes are standardized across sources, multi-author cases are handled uniformly, and each book is classified into a single subject area (e.g., "Fundamentals of Database Systems" classified under "Information Management"). The Prices[] attribute records vendor-specific prices so customers can easily compare across stores, while AvailabilityStatus is normalized to {InStock, OutOfStock}.

The integration system supports the two classes of workloads queries: (1) Find Book by Title & Author (Customer Search – OLTP). (2) Cheapest Vendor for a Book (Customer Purchase – OLTP) (3) User Purchase History with Recommendations (Profile-based – OLTP/OLAP Hybrid) (4) Publisher-Level Sales Trends (Analytics – OLAP), and (5) Price Variability Analysis (Analytics – OLAP). These dual workloads create a need for a hybrid integration system, where some attributes (like prices and availability) are virtualized to ensure real-time freshness via a federated query engine, while others (like book metadata and subject classification) are materialized into a data warehouse using scheduled ETL (Extract, Transform, Load) jobs. The materialized data is populated by a nightly batch process that pulls data from all sources, runs it through data cleansing and entity resolution pipelines, and loads the clean data. The materialized data is then used for efficient analytics and

OLAP queries. View maintenance is a critical challenge, requiring a strategy like incremental updates to keep materialized views consistent without re-computing them from scratch.

An indexing strategy is critical to support these workloads efficiently. B-tree or hash indexes are proposed on ISBN for exact lookups, inverted indexes on Title and Authors for keyword-based searches (often a full-text search engine like Elasticsearch is used for this), and bitmap indexes on categorical attributes such as SubjectArea or AvailabilityStatus. For OLTP queries, a composite index on (Title, Authors[]) can be highly efficient. For OLAP workloads, precomputed materialized aggregate views—such as BooksBySubject, PublisherSales, or UserPurchaseHistory—can be indexed to support fast trend analysis and recommendations. Given storage constraints, a careful selection of which views to materialize must be made, guided by a cost-benefit analysis of storage space versus query performance, which is known as the view selection problem.

✓ 4UT5M. In the SastaBookStore.com system, which architectural component is most responsible for abstracting away the differences between BookStore1's RESTful API and BookStore2's SOAP API?

✓ PTLMG. When integrating vendor-specific *Price* attributes into a global schema, how should different currency values ideally be handled?

Y5CDX. The biggest challenge in mapping Availability/Inventory/InStock fields into a unified global attribute is:

✓ EOZYG. Which of the following is NOT a factor in view selection for materialization?

✓ H9P5. A query like *"Find the cheapest vendor price for ISBN X"* is most efficiently served by:

4RBDZ. For OLAP queries grouping by Publisher, which index structure is most efficient?

BW1XZ. A star schema in SastaBookStore would place which table at the center?

✓ 0266G. Which OLAP operation is used when zooming into "Information Management" books from overall sales?

✓ KFT5G. Which hybrid approach balances OLTP freshness with OLAP efficiency?

✗ J179X. Which of the following causes loss of self-maintainability in views?

✓ UGBIK. Which technique allows automatic refresh of materialized views in SastaBookStore when new transactions are recorded?

2KU0C. Suppose we have 3 candidate materialized views with the following sizes and query cost reduction benefits (in units):

View V1: Size = 20 MB, Benefit = 200	View V2: Size = 30 MB, Benefit = 400
View V3: Size = 20 MB, Benefit = 600	Available storage = 50 MB.

Using the Greedy algorithm (Efficiency = Benefit/Size), which set of views will be selected?

o 89SNT. Suppose we have 2 candidate materialized views with the following sizes and query cost reduction benefits (in units):

View V1: Size = 25 MB, Benefit = 200 View
View V2: Size = 40 MB, Benefit = 320
Available storage = 40 MB.

Using the Greedy algorithm (Efficiency = Benefit/Size), which view is picked first?

✓ YP5XX. Materialized Views (MV) is not self-maintainable if:

✓ OFJYS. If we have 10M records in Bookstore1, and the query distribution is on Title=50%, Author=30% and ISBN=20%. If we create indexes on two attributes, which combination is optimal?

✓ OK35B. If Title index reduces lookup cost from 1000 ms to 50 ms, and Title queries=500/day, total saving?

0 RKY8P. Which index is most beneficial for OLTP workload?

o 6LIIP. For OLAP queries aggregating sales by Year, which index helps most?

✓ V8VMX. Which query is **NOT** answerable directly from Materialized Views (MV) storing only yearly aggregates?

✗ R0JZ8. Which is true about query rewrite engines?

✓ JBL0M. Which mapping technique is best for synonyms like "Price" vs "Unit_Price"?

✓ 026C3. In hybrid integration, which attributes are best materialized?

✓ FXLMC. Suppose TransactionHistory and UserProfiles are materialized, while BookAvailability is virtualized. If 70% of queries hit transactions, 20% hit profiles, and 10% check availability, what is the expected performance gain from this hybrid design?

o K6STM. What is a key challenge of hybrid systems?

✓ BQDX5. Which is the most important decision factor in view selection?

✓ KC9SB. What is the primary benefit of materializing data for Publisher-Level Sales Trends analysis instead of virtualizing it?

o QBWIC. Which workload type in SastaBookStore.com is most dependent on the pre-computed materialized views stored in the data warehouse?

✓ FJ75Z. The use of a global unified schema in SastaBookStore.com's system design is a core tenet of which information integration paradigm?

LG208. The use of an automated tool to suggest that Pub-Year from BookStore1 and Year-of-Publication from BookStore2 refer to the same concept based on their name similarity is an example of what?

RBV60. In the SastaBookStore.com system, resolving the different representations of "Publisher" (Publisher, PublisherName, Pub) requires addressing which type of schema conflict?