Syllabus

Applied Database System Engineering (ADSE)

Theory and Lab

Specialization: *Database Engineering and OS*
Pre-Requisite Courses Completion: *RDBMS, Operating Systems*
Version: *3.1, 11-Dec-2024*

Course Objectives: *This course intends to introduce topics related to Distributed Database Engineering and its applications in various business domains together with knowledge on how to develop products over them. Also, in depth knowledge of kernel processes that interface with database engines.*

Course Outcome: *This course shall equip the computer science and information technology students to skills required for the development of products related to Distributed Databases. This subject introduces deep knowledge in structure of Distributed Databases to the record level and impart skill to understand challenges in Distributed Systems and to derive possible solutions to these challenges. And skills that is required to support changes that to be done at kernel level for effective implementation of database applications.*

Content of the Course

Theory (5 Units x 8 hrs = 40 hrs)

UNIT 1 – DB Engine and Kernel Interface (8 hrs)

DB engine architecture and functions - Kernel managements - Kernel's device interface - Memory and file managements of kernel - Kernel's hooks for optimization of data transfers

UNIT 2 – Advanced IPC Mechanisms for Database Operations (8 hrs)

Shared memory, mmap and shmget - shared memory synchronization techniques - Message queues - Netlink sockets for IPC - Futexes(Fast Userspace Mutexes) - Remote Direct Memory Access (RDMA) communication - Remote Procedure Calls (gRPC) framework - Desktop Bus(D-Bus)

UNIT 3 – Algorithms - Memory Management and DB Engines (8 hrs)

Page replacement algorithms, LRU, LFU and CLOCK - Memory fragmentation and defragmentation techniques - Paging, Memory Mapping, TLB - NUMA - aware memory management - Huge Pages, Transparent Huge Pages (THP), regressions in THP - Sorting & Aggregations Algorithms - Memory allocation and caching strategies in DB engines

UNIT 4 – DB Performance Measurement Strategies and Techniques (8 hrs)

Performance metrics, Latency, throughput, cache hits/misses, query execution times - Query optimization and indexing - Benchmarking tools like pgbench and sysbench - Proofing tools like eBPF, Single Root I/O Virtualization (SR-IOV) - Storage I/O performance, Monitoring tools like Prometheus and Grafana - Query Planning & Optimization - Concurrency Control Theory.

UNIT 5 - Distributed Database Architecture and its Data Structures (8 hrs)

Architectural models for distributed DBMSs - Client/Server systems - Peer-to-Peer systems - Multidatabase systems - Security Implications of Memory Deduplication in a Virtualized Environment - B-Tree Indexes - Hash Indexes - GiST (Generalized Search Tree) Indexes - Bitmap Indexes - Column-Store Indexes - AgentsGraph - TimescaleDB - JSON/XML

LAB EXERCISES and CASE STUDY (8 x 5 hrs = 40 hrs)

Lab Exercise 1: To develop a client-server application program with database access.

Lab Exercise 2: To develop a C program which uses shared memory management of kernel.

Lab Exercise 3: To write a C program to understand how the bitmap indexes can optimize complex AND and OR query combinations and discuss the scenarios in which bitmap indexing is most effective, especially in product sales and marketing applications.

Lab Exercise 4: To write a C program to detect fraud during a financial transaction.

Lab Exercise 5: To write a C program to understand how a custom page replacement algorithm works.

Lab Exercise 6: To write a C program to understand how churn prediction works.

Lab Exercise 7: To write a C program to decide on friendship factor for a given user in a social media.

Lab Exercise 8: Develop a script that automates database performance monitoring.

Text Books:

1. Principles of Distributed Database Systems, M. Tamer Özsu, University of Waterloo and Patrick Valduriez University of Montpellier, 4th Edition, Springer Nature Switzerland AG 2020, Hardcover ISBN 978-3-030-26252-5.

2. Database Reliability Engineering: Designing and Operating Resilient Database Systems, Laine Campbell and Charity Majors, O'Reilly Media; 1st edition, 2017, ISBN-13, 978-1491925942.

3. Linux Kernel Development, Third Edition, Robert Love, 2010, Addison-Wesley Professional, ISBN-10 8131758184.

Reference Books:

1. PostgreSQL Internals, Egor Rogov, Postgres Professional, 2022–2023, ISBN 978-5-6045970-4-0.

2. Database System Concepts, Seventh Edition, Avi Silberschatz, Henry F. Korth, S. Sudarshan, 2011, McGraw-Hill, ISBN 978-0-07-352332-3

3. Designing Data-Intensive Applications, Martin Kleppmann, O'Reilly Media, Inc., 2017, ISBN: 9781491903100.

4. Database Internals: A Deep Dive into How Distributed Data Systems Work, Alex Petrov, First Edition, O'Reilly, ISBN-13 978-1492040347, 2019
