

DATA ENGINEERING AND ANALYTICS

Subject Code: CS35211CS	Subject Type: DC	Evaluation Scheme: MSE (30), TA (20), ESE (50)	L-T-P: 2-0-2 (3)
--------------------------------	-------------------------	--	-------------------------

Pre-Requisites: Database Management Systems, SQL Programming, Data Structures, Basic Programming Knowledge (Python/Java), and Fundamentals of Distributed Systems

Course Objective: To understand the evolution of database systems from relational databases to modern distributed data platforms. To study advanced database architectures designed for Big Data and large-scale distributed systems. To explore different NoSQL database models including document, column-family, key-value, and graph databases.

Course Outcomes: At the end of the course, the student will be able to

CO	Course Outcomes
CO1	Analyze modern database architectures and explain the evolution from traditional RDBMS to NoSQL and distributed databases.
CO2	Design data models using document, column-family, key-value, and graph database systems.
CO3	Evaluate distributed database concepts such as sharding, replication, CAP theorem, and consistency models.
CO4	Implement and analyze modern database technologies including MongoDB, Cassandra, HBase, and Neo4j.
CO5	Assess emerging database systems such as NewSQL, cloud-native databases, and data-intensive architectures.

Syllabus:

Unit 1: Evolution of Database Systems and Big Data Foundations: Data Management Concepts, Structured Query Language (SQL), Data Types and Constraints in MySQL, SQL for Data Definition (DDL), SQL for Data Manipulation (DML), SQL for Data Query (DQL), Data Update and Deletion, Transactions and ACID Properties. Early Database Systems, Hierarchical and Network Models, Relational Database Systems, First Database Revolution, Second Database Revolution, Third Database Revolution, Limitations of Relational Databases. Big Data Revolution, Characteristics of Big Data (Volume, Velocity, Variety, Veracity, Value), Introduction to Hadoop Ecosystem, Hadoop Distributed File System (HDFS), MapReduce Overview, Hive, Pig, Return of SQL in Big Data Systems.

Unit 2: Distributed Databases and NoSQL Foundations: Distributed Database Architecture, Distributed System Characteristics, Data Partitioning, Horizontal and Vertical Partitioning, Sharding Techniques, Replication Strategies. Consistency Models (Strong Consistency, Eventual Consistency, Causal Consistency), ACID vs BASE Properties, CAP Theorem, Distributed Transactions, Two-Phase Commit (2PC). NoSQL Databases Overview, Types of NoSQL Databases (Key-Value, Document, Column-Family, Graph Databases), NoSQL Data Models

Unit 3: Document and Key-Value Databases: Document Data Model, Documents and Collections, JSON Data Representation, Embedded Data Models, Normalized Data Models. MongoDB Architecture, MongoDB Data Manipulation Language, CRUD Operations, indexing in MongoDB, MongoDB Replication using Replica Sets, MongoDB Sharding. Key-Value Databases, Redis Architecture, Key-Value Database Use Cases.

Unit 4: Column-Family Databases: Column-Family Data Model, Wide Column Databases, Databases and Tables, Columns, Types and Keys. Apache Cassandra Architecture, Cassandra Query Language (CQL), Keyspaces and Column Families, Data Partitioning in Cassandra, Replication Strategies, Cluster Management, Fault Tolerance. Hadoop HBase Architecture, HBase Data Model, Comparison of Cassandra and HBase, Large-scale Distributed Storage Systems.

Unit 5: Graph Databases and Emerging Database Technologies: Graph Data Model, Nodes, Edges, and Properties, Graph Database Use Cases. Neo4j Architecture, Neo4j Deployment (Standalone and Cluster), Cypher Query Language, Querying Graph Data, Applications in Social Networks, Recommendation Systems, Fraud Detection Systems, Knowledge Graphs. NewSQL Databases, Google Spanner Architecture (Overview), Distributed SQL Databases, Cloud Databases, Multi-model Databases, Database-as-a-Service (DBaaS), Data Lakes and Data Warehouses, Modern Data-Intensive Architectures.

Learning Resources:

Text Books:

1. Sadalage, P. & Fowler, M. NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence. Pearson Education, 2012.
2. Redmond, E. & Wilson, J. Seven Databases in Seven Weeks: A Guide to Modern Databases and the NoSQL Movement. Pragmatic Programmers, 2012.
3. Dan Sullivan. NoSQL for Mere Mortals. Addison-Wesley Professional, 2015.
4. Guy Harrison. Next-Generation Databases. Apress, 2016.
5. Martin Kleppmann. Designing Data-Intensive Applications. O'Reilly Media, 2017.