**BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI**

**WORK INTEGRATED LEARNING PROGRAMMES**

**Part A: Content Design**

|  |  |
| --- | --- |
| **Course Title** | Distributed Data Systems |
| **Course No(s)** | SS ZG554 |
| **Credit Units** |  |
| **Credit Model** |  |
| **Content Authors** | Anil Kumar G |

**Course Objectives**

|  |  |
| --- | --- |
| **No** | **Course Objectives** |
| **CO1** | This field covers all aspects of computing and information access across multiple processing elements connected by any form of communication network, either local area, or wide area |
| **CO2** | There has been a steady growth in the development of contemporary applications that demonstrate their efficacy by connecting millions of users/applications/machines across the globe without relying on a traditional client-server approach. |
| **CO3** | The general computing trend is to leverage shared resources and massive amounts of data over the Internet. This course aims to provide an understanding of theory and systems aspects of distributed data |

**Text Book(s)**

|  |  |
| --- | --- |
| **T1** | M. Tamer Özsu • Patrick Valduriez Principles of Distributed Database Systems Third Edition |
| **T2** | Distributed Operating Systems: Concepts And Design By Pradeep K. Sinha |

**Reference Book(s)**

|  |  |
| --- | --- |
| **R1** | “Storage Networks Explained” – by Ulf Troppens, Wolfgang Muller-Freidt, Rainer Wolafka, IBM Storage Software Development, Germany. Publishers: Wiley |

**On-Line Resources**

**HBase**

<https://hbase.apache.org>

<http://www.tutorialspoint.com/hbase/>

**MapReduce**

<https://www-01.ibm.com/software/data/infosphere/hadoop/mapreduce/>

<http://www.tutorialspoint.com/hadoop/hadoop_mapreduce.htm>

**SAN**

<http://searchstorage.techtarget.com/definition/storage-area-network-SAN>

<http://www.snia.org/education/storage_networking_primer/san/what_san>

**NAS**

<http://searchstorage.techtarget.com/definition/network-attached-storage>

<http://www.webopedia.com/TERM/N/network-attached_storage.html>

**Content Structure**

1. Distributes Data Storage Technology
   1. Server-centric IT architecture and its limitations
   2. Storage-centric IT architecture and its advantages
   3. Architecture of intelligent disk subsystems
   4. Hard disks and internal i/o channels and JBOD
   5. Storage virtualization using RAID
   6. Introduction to NAS, SAN and DAS
2. Distributed File Systems & Security
   1. File Models & Accessing models
   2. File sharing Semantics
   3. File Caching
   4. File Replication
   5. Fault Tolerance
   6. File System Security
3. Distributed Databases
   1. Distributed DBMS
   2. Distributed DBMS Architecture
   3. Distributed Data Sources
   4. Distributed Design Issues
4. Distributed Database Design Issues & Integration
   1. Distributed Design Issues
   2. Top-Down Design Process
   3. Fragmentation
   4. Bottom-Up Design Methodology
   5. Allocation
   6. Schema Matching
   7. Schema Integration
   8. Schema Mapping
   9. Data Cleaning
5. Data and Access Control
   1. Database Security
   2. Discretionary Access Control
   3. Multilevel Access Control
   4. Distributed Access Control
   5. View Management
   6. Views in Centralized DBMSs
   7. Views in Distributed DBMSs
   8. Maintenance of Materialized Views
6. Data Replication
   1. Consistency of Replicated Databases
   2. Update Management Strategies
   3. Replication Protocols
   4. Replication and failures
   5. Replication Mediator Service
7. Parallel Database Systems
   1. Parallel Database System Architectures
   2. Parallel Data Placement
   3. Load Balancing
   4. Database Clusters
8. Web Data Management
   1. Web Search
   2. Web Crawling
   3. Indexing
   4. Ranking and Link Analysis
   5. Keyword Search
   6. Web Querying
   7. Semi-structured Data Approach
   8. Web Query Language Approach
   9. Question Answering
   10. Searching and Querying the Hidden Web
9. Hadoop & Big Data
   1. Introduction
   2. HDFS Operations
   3. Map Reduce
   4. Big Data Overview
   5. HDFS Commands
   6. Setup steps
   7. Multi Node Cluster
   8. Big Data Solutions

**Learning Outcomes:**

|  |  |
| --- | --- |
| **No** | **Learning Outcomes** |
| **LO1** | Understanding about Distributed structures |
| **LO2** | Understanding of Distributed Storage systems and the technologies used to implement |
| **LO3** | Understanding of Distributed databases architecture |
| **LO4** | Understanding of Parallel databases architecture and systems |
| **LO5** | Understanding Hadoop environment and Big Data |

**Part B: Learning Plan**

|  |  |
| --- | --- |
| **Academic Term** | SECOND SEMESTER 2015-2016 |
| **Course Title** | Distributed Data Systems |
| **Course No** | SS ZG554 |
| **Lead Instructor** | Anil Kumar G |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Session** | **Title** | **Description** | **Exercise** | **References** |
| 1 | Distributes Data Storage Technology | * Server-centric & Storage-centric IT architecture with its limitations and advantages * Architecture of intelligent disk subsystems * Hard disks and internal i/o channels and JBOD * Storage virtualization using RAID * Introduction to NAS, SAN and DAS | Understanding RAID Levels 4 & 5  R1 : Page 535 & 536 | R1 - Ch.1  R1 - Ch.2  R1 - Ch.4 |
| 2 | Distributed File Systems & Security | * File Models & Accessing models * File sharing Semantics * File Caching * File Replication * Fault Tolerance * File System Security | - | T2 – Ch.9 |
| 3 | Distributed Databases | * Distributed DBMS * Distributed DBMS Architecture * Distributed Data Sources * Distributed Design Issues | - | T1 – Ch.1 |
| 4 | Distributed Database Design | * Top-Down Design Process * Distributed Design Issues * Fragmentation * Allocation | Solve  T1 : Problem 3.1 & 3.2 Page 126 | T1 – Ch.3 |
| 5 | Distributed Database Integration | * Bottom-Up Design Methodology * Schema Matching * Schema Integration * Schema Mapping * Data Cleaning | Solve  T1 : Problem 4.4 Page 161 | T1 – Ch.4 |
| 6 | Data and Access Control | * View Management * Views in Centralized DBMSs * Views in Distributed DBMSs * Maintenance of Materialized Views | - | T1 – Ch.5 |
| 7 | Data and Access Control | * Database Security * Discretionary Access Control * Multilevel Access Control * Distributed Access Control | Solve  T1 : Problem 5.1 Page 202 | T1 – Ch.5 |
| 8 | Mid Term Review | Review of Sessions 1 to 7 | - | - |
| 9 | Data Replication | * Consistency of Replicated Databases * Update Management Strategies * Replication Protocols | - | T1 – Ch.13 |
| 10 | Data Replication | * Replication and failures * Replication Mediator Service | Solve  T1 : Problem 13.2 Page 493 | T1 – Ch.13 |
| 11 | Parallel Database Systems | * Parallel Database System Architectures * Parallel Data Placement * Load Balancing * Database Clusters | Solve  T1 : Problem 14.15 Page 550 | T1 – Ch.14 |
| 12 | Web Data Management | * Web Search * Web Crawling * Indexing * Ranking and Link Analysis   Keyword Search | - | T1 – Ch 17 |
| 13 | Web Data Management | * Web Querying * Semi-structured Data Approach * Web Query Language Approach * Question Answering * Searching and Querying the Hidden Web | Solve  T1 : Problem 17.1 Page 719 | T1 – Ch.17 |
| 14 | Hadoop & Big Data | * Introduction * HDFS Operations * HDFS Commands * Map Reduce | - | Online Resources |
| 15 | Hadoop & Big Data | * Multi Node Cluster * Big Data Overview * Big Data Solutions | - | Online Resources |
| 16 | Compre Review | Review of Sessions 9 to15 | - | - |

**Evaluation Scheme**:

Legend: EC = Evaluation Component; AN = After Noon Session; FN = Fore Noon Session

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| No | Name | Type | Duration | Weight | Day, Date, Session, Time |
| EC-1 | Quiz-I/ Assignment-I | Online | - | 5% | September 1-10, 2016 |
|  | Quiz-II | Online |  | 5% | October 1-10, 2016 |
|  | Quiz-III/ Assignment-II | Online |  | 5% | October 20-30, 2016 |
| EC-2 | Mid-Semester Test | Closed Book | 2 hours | 35% | 25/09/2016 (AN) 2 PM TO 4 PM |
| EC-3 | Comprehensive Exam | Open Book | 3 hours | 50% | 06/11/2016 (AN) 2 PM TO 5 PM |

Syllabus for Mid-Semester Test (Closed Book): Topics in Session Nos. 1 TO 8

Syllabus for Comprehensive Exam (Open Book): All topics (Session Nos. 1 to 16)

**Important links and information:**

Elearn portal: https://elearn.bits-pilani.ac.in

Students are expected to visit the Elearn portal on a regular basis and stay up to date with the latest announcements and deadlines.

Contact sessions:Students should attend the online lectures as per the schedule provided on the Elearn portal.

Evaluation Guidelines:

1. EC-1 consists of either two Assignments or three Quizzes. Students will attempt them through the course pages on the Elearn portal. Announcements will be made on the portal, in a timely manner.
2. For Closed Book tests: No books or reference material of any kind will be permitted.
3. For Open Book exams: Use of books and any printed / written reference material (filed or bound) is permitted. However, loose sheets of paper will not be allowed. Use of calculators is permitted in all exams. Laptops/Mobiles of any kind are not allowed. Exchange of any material is not allowed.
4. If a student is unable to appear for the Regular Test/Exam due to genuine exigencies, the student should follow the procedure to apply for the Make-Up Test/Exam which will be made available on the Elearn portal. The Make-Up Test/Exam will be conducted only at selected exam centres on the dates to be announced later.

It shall be the responsibility of the individual student to be regular in maintaining the self study schedule as given in the course handout, attend the online lectures, and take all the prescribed evaluation components such as Assignment/Quiz, Mid-Semester Test and Comprehensive Exam according to the evaluation scheme provided in the handout.