# BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI

## WORK INTEGRATED LEARNING PROGRAMMES

## COURSE HANDOUT

## Part A: Content Design

| | |
|---|---|
| **Course Title** | Big Data Systems |
| **Course No(s)** | SE* ZG522 |
| **Credit Units** | 5 |
| **Course Author** | Prof. Shan Balasubramaniam |
| **Version No** | 3.5 |
| **Last Revised By** | Janardhanan PS |
| **Date** | 07 November 2023 |

### Course Description

The course introduces the students to systems for data analytics with particular emphasis on storage and processing of Big Data. It introduces computing models for distributed processing for scalability and fault-tolerance. It covers frameworks and tools for ingestion and batch processing of data stored on distributed file systems, in-memory distributed processing and stream processing for real time analytics. NoSQL databases are covered in detail. The Big Data Systems course is designed to impart an in-depth knowledge of Big Data processing using NoSQL databases, Hadoop and Spark. This course will help the students to master essential skills on NoSQL databases, Hadoop eco-system products, Apache Spark framework including Spark SQL, Spark Streaming and Machine learning programming. Amazon's storage and database services are used as exemplar platforms on Cloud.

### Course Objectives

| | |
|---|---|
| **CO1** | Enable students to understand requirements for and constraints in storing and processing Big Data |
| **CO2** | Enable students to leverage commodity infrastructure (such as scale-out clusters, distributed data stores, and the cloud) and the appropriate platforms and services for storing and processing Big Data. |
| **CO3** | Enable students to implement solutions for big data processing |
| **CO4** | Enable students to develop a working knowledge of stream processing |

### Text Book(s)

| | |
|---|---|
| T1 | Seema Acharya and Subhashini Chellappan. *Big Data and Analytics*. Wiley India Pvt. Ltd. Second Edition |
| T2 | Raj Kamal and Preeti Saxena, *Big Data Analytics*. McGraw Hill Education (India) Pvt.Ltd |

**Reference Book(s) & other resources**

| | |
|---|---|
| R1 | DT Editorial Services. *Big Data - Black Book*. DreamTech. Press. 2016 |
| R2 | Kai Hwang, Jack Dongarra, and Geoffrey C. Fox. *Distributed and Cloud Computing: From Parallel Processing to the Internet of Things*. Morgan Kauffman 2011 |
| R3 | *Additional Reading* (as assigned for specific topics) |

**Learning Outcomes:**

| **No** | **Learning Outcomes** |
|---|---|
| **LO1** | A comprehensive understanding of the Big Data ecosystem and along with the typical technologies involved. |
| **LO2** | Apply concepts from distributed computing and use the Hadoop/Map-reduce framework and for solving typical big data problems. |
| **LO3** | Identify and use appropriate storage / database platforms for Big data storage along with appropriate querying mechanisms / interfaces for retrieval. |
| **LO4** | Use in-memory processing and stream processing techniques for building Big Data systems. |

# Modular Structure

| Module # | Name of Module | Contact Sessions |
|---|---|---|
| 1 | **Data Engineering** | 1 - 2 |
| 2 | **Big Data Analytics** | 3 - 5 |
| 3 | **Hadoop Ecosystem** | 6 - 9 |
| 4 | **NoSQL Databases** | 10 |
| 5 | **Spark for Big Data Processing** | 11 – 14 |
| 6 | **Big Data Storage and Processing on Cloud** | 15 - 16 |

# Part B: Contact Session Plan

| | |
|---|---|
| **Academic Term** | S1-23  Sem 3 |
| **Course Title** | Big Data Systems |
| **Course No** | DSECL ZG522 |
| **Lead Instructor** | Janardhanan PS |

| Session # | Contact Hours(#) | List of Topic / Title | Text/Ref Book/external resource |
|---|---|---|---|
| 1 | 1 | **Different Types of Data and Storage for Data**:<br>Structured Data (Relational Databases) , Semi-structured data (Object Stores), and Unstructured Data (File systems)<br>**What is Big Data?**<br>Characteristics of Big Data.<br>Systems perspective - Processing: In-memory vs. (from) secondary storage vs. (over the) network | T1 Ch. 1<br>T1 Ch.2<br>R2 Sec 1.2.3 |
|  | 2 | **Locality of Reference**: Principle, examples<br>**Impact of Latency**: Algorithms and data structures that leverage locality, data organization on disk for better locality | Class Slides |
| 2 | 3 | **Parallel and Distributed Processing**: Motivation (Size of data and complexity of processing); Storing data in parallel and distributed systems: Shared Memory vs. Message Passing; Strategies for data access: Partition, Replication, and Messaging. | R2 Sec. 1.2, 1.3.4, and 1.4.1 |
|  | 4 | **Memory Hierarchy in Distributed Systems:** In-node vs. over the network latencies, Locality, Communication Cost.<br>**Distributed Systems:** Motivation (size, scalability, cost-benefit), Client-Server vs. Peer-to-Peer models, Cluster Computing: Components and Architecture | Class Slides<br>R2 Sec. 2.1 to 2.3 |
| 3 | 5 | **Big Data Analytics**: Requirements, constraints, approaches, and technologies. | T1 Sec. 3.1 to 3.11; R1 Ch. 3 and Ch. 6 |
|  | 6 | **Big Data Systems – Characteristics:** Failures; Reliability and Availability; Consistency – Notions of Consistency. | T1 Ch. 4<br>AR |
| 4 | 7 | **CAP Theorem and implications for Big data Analytics** | T1 Sec. 3.12 and 3.13; AR |
|  | 8 | **Big Data Lifecycle:** Data Acquisition, Data Extraction – Validation and Cleaning, Data Loading, Data Transformation, Data Analysis and Visualization. Case study – Big data application | T1 Sec. 2.9 to 2.12; R1 Ch. 6 and Ch. 7 |
| 5 | 9-10 | **Distributed Computing - Design Strategy:** Divide-and-conquer for Parallel / Distributed Systems - Basic scenarios and Implications.<br>**Programming Patterns:** Data-parallel programs and *map* as a construct; Tree-parallelism, and *reduce* as a construct; Map-reduce model: Examples (of map, reduce, | AR |

| | | | |
|---|---|---|---|
| | | map-reduce combinations, and Iterative map-reduce) | |
| 6 | 11-12 | **Hadoop:** Introduction, Architecture, and Map-reduce Programming on Hadoop | T1 Sec. 5.1 and 5.2, Sec. 5.7, Sec. 5.11, and Ch. 8; T2 Sec.2.1, R1 Ch. 5 and Ch. 9; R2 Sec. 1.4.3 and 6.2.2; AR |
| 7 | 13-14 | **Hadoop**: Hadoop Distributed File System (HDFS), Scheduling in Hadoop (using YARN). Example – Hadoop application. | T1 5.10 and 5.12; R1 Ch. 4 (sections on HDFS and Yarn) and Ch. 11 of R1 |
| 8 | 15-16 | **Hadoop Ecosystem:** Databases and Querying (HBase, Pig, and Hive) | T1 Sec. 5.13;<br>R1 Ch. 4 (sections on HBase, Hive, and Pig) and Ch. 5 (section on HBase) |
| | | **Mid Semester Examination** | |
| 9 | 17-18 | **Hadoop Ecosystem:** Integration and coordination (Sqoop, Flume, Zookeeper & Oozie) | T1 Sec. 5.13;<br>R1 Ch. 4 (sections on Sqoop, Flume, Zookeeper & Oozie) |
| 10 | 19-20 | **NoSQL databases:** Introduction, Architecture, Querying, Variants, Case Study. | T1 Sec. 4.1, Ch. 6, and Ch. 7, T2 Ch.3. |
| 11 | 25 | **Spark:** Introduction, Architecture and Features | T2 Ch.5 |
| | 26 | **Programming on Spark:** Resilient Distributed Datasets, Transformation, Spark SQL, Examples | T2 Sec.5.3, AR (Apache Spark docs.) |
| 12 | 27-28 | **Machine Learning (on Spark):** Regression, Classification, Collaborative Filtering, and Clustering. | T1 Ch.12, T2 Ch.6, AR (Apache Spark docs.) |
| 13 | 29-30 | **Streaming:** Stream Processing – Motivation, Examples, Constraints, and Approaches. | AR |
| 14 | 31-32 | **Streaming on Spark:** Architecture of Spark Streaming, Stream Processing Model, Example. | T2 Ch.7 |
| 15 | 21 | **Cloud Computing:** A brief overview: Motivation, Structure and Components; Characteristics and advantages – Elasticity, Dynamic provisioning, Multi-tenancy. Services on the cloud. | AR |
| | 22 | **Storage as a Service:** Forms of storage on the cloud, Cloud managed NoSQL databases. | AR |

|  |  |  |  |
|---|---|---|---|
| 16 | 23 | **Amazon's storage services**: block storage, file system, and database;  EBS, SimpleDB, S3 | AR (*sourced from Amazon*) |
|  | 24 | Case study – Amazon DynamoDB (Access/Querying model, Database architecture and applications on the cloud). | - |

**Topics for experiential learning:**

| Topic No. | Selected Topics in Syllabus for experiential learning |
|---|---|
| 1 | • Exercises on Distributed Systems – Hadoop;<br>• Exercises using Map-reduce model:  Standard patterns in map reduce models. |
| 2 | • Exercises on NoSQL;<br>• Exercises with NoSQL database – Simple CRUD operations and Failure / Consistency tests;<br>• Cassandra - Consistency levels, CRUD operations, Schema on read<br>• MongoDB<br>• Neo4j Graph database - Relationships and queries<br>• HBase queries |
| 3 | • Exercises with Pig queries to perform Map-reduce job and understand how to build queries and underlying principles;<br>• Exercises on creating Hive databases and operations on Hive, exploring built in functions, partitioning, data analysis |
| 4 | • Exercises on Spark to demonstrate RDD, and operations such as Map, FlatMap, Filter, PairRDD;<br>• Typical Spark Programming idioms such as : Selecting Top N, Sorting, and Joins;<br>• Exercises on DataFrames, Datasets, and Spark SQL<br>• Spark Streaming - Sample Streams, Structured Streaming, Windowed Streaming |
| 5 | • Exercises using Spark MLlib: Regression, Classification, Collaborative Filtering, Clustering |
| 6 | • Exercises on Analytics on the Cloud – using AWS S3, AWS EMR, AWS data stores / databases, Querying with DynamoDB. |

[**Note**: *Most of the topics for experiential learning will be covered by hands on demonstrations and/or participatory lab sessions operated remotely. Rest of them will be assigned as homework and may be included for evaluation*]

**Webinars:**

| No. | On Selected Topics in the Syllabus |
|---|---|
| 1 | As per Programme Calendar |
| 2 | As per Programme Calendar |
| 3 | As per Programme Calendar |
| 4 | As per Programme Calendar |

## Evaluation Scheme

Legend: EC = Evaluation Component

| No | Name | Type | Duration | Weight | Day, Date, Session, Time |
|---|---|---|---|---|---|
| EC-1 | Assignment I | Take-home: Configuration of Bigdata platforms, Programming and querying to solve analytics problems | | ( 12 +13 ) = 25 % | As per Programme Calendar |
| | Assignment  II | | | | As per Programme Calendar |
| | Quiz (Best of 2 quizzes) | Online, at scheduled time | | 5% | As per Programme Calendar |
| EC-2 | Mid-Semester Exam | Closed Book | 2.0 hours | 30% | As per Programme Calendar |
| EC-3 | Comprehensive Exam | Open Book | 2.5 hours | 40% | As per Programme Calendar |

## Important Information

Syllabus for Mid-Semester Test (Open Book): Topics in Weeks 1-8
Syllabus for Comprehensive Exam (Open Book): All topics given in plan of study

Evaluation Guidelines:

1. EC-1 consists of two Assignments and one quiz (best out of two). Announcements regarding the same will be made in a timely manner.
2. For Closed Book tests: No books or reference material of any kind will be permitted. Laptops/Mobiles of any kind are not allowed. Exchange of any material is not allowed.
3. For Open Book exams: Use of prescribed and reference textbooks, in original (not photocopies) is permitted. Class notes/slides as reference material in filed or bound form is permitted. All other additional reading materials in filed / bound form are also 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. The genuineness of the reason for absence in the Regular Exam shall be assessed prior to giving permission to appear for the Make-up Exam. 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 lectures, and take all the prescribed evaluation components such as Assignment/Quiz, Mid-Semester Test and Comprehensive Exam according to the evaluation scheme.