Maharashtra State Board of Technical Education, Mumbai

LABORATORY PRACTICAL PLANNING

K2-A

Academic Year: 2025-26 **Date: 12/12/2025**
Institute Name & Code: K. K. Wagh Polytechnic, Nashik-3 (0078)
Program & Code: Artificial Intelligence & Machine Learning (AN) **CourseCode&Abbr.:**316318(BDA)
Course Name: Big Data Analytics **Name of Faculty**: Mrs. J.S.Mahajan
Class: TYAN **Course Index:** 602 **Semester**: VI[rd] **Scheme**: 'K' **Total Hrs:** 60

- ## Course Outcomes (COs):

By learning course Big Data Analytics (BDA-316318) Third Year students will be able to:

CO602.1- Illustrate different phases of big data with respect to real world application.
CO602.2- Demonstrate the use of Hadoop core components for big data processing.
CO602.3- Apply NoSQL database concepts and architecture patterns to manage big data.
CO602.4- Use Hive and Pig for data processing and transformation within big data environments.
CO602.5- Use Spark to process and analyze big data in real-time or archives.

- ## Teaching-Learning and Assessment Scheme:

Course Code	Course Title	Abbr	Course Category/s	Learning Scheme					Credits	Assessment Scheme											
				Actual Contact Hrs./Week			SLH	NLH		Paper Duration	Theory				Based on LL & TL				Based on SL		Total Marks
															Practical						
				CL	TL	LL					FA-TH	SA-TH	Total		FA-PR		SA-PR		SLA		
											Max	Max	Max	Min	Max	Min	Max	Min	Max	Min	
316318	BIG DATA ANALYTICS	BDA	DSC	3	-	4	1	8	4	3	30	70	100	40	25	20	-	-	25	10	150

Abbreviations: CL CL- ClassRoom Learning , TL- Tutorial Learning, LL-Laboratory Learning, SLH-Self Learning Hours, NLH-Notional Learning Hours, FA - Formative Assessment, SA -Summative assessment, SLA - Self Learning Assessment

- ## Laboratory Learning Outcome (LLO):

LLO 1.1 Identify Big Data use cases and explain the analytics process applied.
LLO 2.1 Setup a Hadoop ecosystem on a local cluster.
LLO 3.1 Configure Hadoop Ecosystem on Local Cluster.
LLO 4.1 Setup multi-node hadoop cluster.
LLO 5.1 Configure multi-node hadoop cluster.
LLO 6.1 Use resource utilization and performance tools in Hadoop Cluster.
LLO 7.1 Execute file operations using HDFS commands.
LLO 8.1 Execute basic backup and restore operations for Big Data stored in HDFS.
LLO 9.1 Develop a MapReduce program.
LLO 10.1 Execute a data processing workflow with MapReduce for CSV files.
LLO 11.1 Setup a MongoDB NoSQL database with collections.
LLO 12.1 Create a schema for unstructured data in MongoDB using any dataset.
LLO 13.1 Run basic aggregation queries on Unstructured dataset in MongoDB.
LLO 14.1 Apply basic operations in MongoDB to observe the behavior of CAP theorem properties.
LLO 15.1 Install Hive within a Hadoop environment.

LLO 16.1 Execute after writing complex data queries using HiveQL.
LLO 17.1 Run after writing Pig scripts for data transformation.
LLO 18.1 Use after writing User Defined Functions (UDF) in Pig.
LLO 19.1 Install Spark on a cluster and verify installation.
LLO 20.1 Perform data transformations with RDDs in Spark.
LLO 21.1 Execute an end-to-end ETL process using Spark.
LLO 22.1 Load structured data and create temporary views in Spark SQL.
LLO 23.1 Run after writing basic SQL queries on datasets using Spark SQL.
LLO 24.1 Stream real- time text and count words using Spark streaming.
LLO 25.1 Save after extracting streaming data into persistent storage using Spark Streaming.
LLO 26.1 Run after creating a regression model on a large dataset in Spark.
LLO 27.1 Apply the K-Means clustering algorithm in Spark MLlib.
LLO 28.1 Visualize clustering results of a dataset processed with Spark MLlib.
LLO 29.1 Classify a dataset using the Decision Tree algorithm in Spark MLlib.
LLO 30.1 Display classification outputs from Decision Tree results.

- **Laboratory Plan:**

Sr. No.	Course Outcomes	LLO	Name of Practical	Planned Date	Performance Date	Remark	Related Self Learning (if any)
1	CO602.1	1.1	Conduct a Case Study on Big Data and Big Data Analysis	A:17/12/2025 B:15/12/2025 C:16/12/2025			
2	CO602.1	2.1	Install Hadoop Ecosystem on Local Cluster	A:20/12/2025 B:18/12/2025 C:19/12/2025			
3	CO602.1	3.1	Configure Hadoop Ecosystem on Local Cluster 1. Configure core-site.xml,hdfs-site.xml, mapred-site.xml 2. Start HDFS daemons 3. Upload any dataset to HDFS	A:24/12/2025 B:22/12/2025 C:23/12/2025			
4	CO602.1	4.1	Install Hadoop on multiple nodes	A:27/12/2025 B:01/01/2026 C:26/12/2025			
5	CO602.1	5.1	Configure Multi-Node Hadoop Cluster 1. Configure core-site.xml and hdfs-site.xml for multi-node HDFS setup 2. Start NameNode, DataNode daemons across nodes	A:03/01/2026 B:05/01/2026 C:02/01/2026			

			3. Upload any dataset to HDFS and verify distributed storage				
6	CO602.1	6.1	Use Monitoring tools to Observe Cluster Resources 1. Access Hadoop monitoring interfaces (CLI or web UI) 2. Locate resource usage indicators (CPU, memory, disk) 3. Run basic monitoring commands 4. Record resource values during cluster operation	A:07/01/2026 B:08/01/2026 C:06/01/2026			
7	CO602.1	7.1	Perform Basic File Operation using HDFS 1. Create directories and Files in HDFS 2. Perform read, write, update and delete operations 3. Set directory permissions 4. Verify file replication	A:10/01/2026 B:12/01/2026 C:09/01/2026			
8	CO602.1	8.1	Perform Backup and Restore of Datasets in HDFS 1. Load any datasets in HDFS 2. Copy datasets to a backup directory 3. Simulate accidental data removal 4. Restore the dataset from the backup	A:14/01/2026 B:15/01/2026 C:13/01/2026			
9	CO602.1	9.1	Execute WordCount MapReduce Program 1. Load any text file into HDFS 2. Develop a Word-Count program in Java 3. Compile, execute, and validate output	A:17/01/2026 B:19/01/2026 C:16/01/2026			
10	CO602.1	10.1	Process Large CSV Dataset Using MapReduce 1. Load a CSV dataset into HDFS 2. Run a MapReduce job to extract specific fields 3. Aggregate data using the Reducer	A:21/01/2026 B:22/01/2026 C:20/01/2026			

11	CO602.1	11.1	Install NoSQL Database (MongoDB) and Create Collections	A:24/01/2026 B:02/02/2026 C:23/01/2026			
12	CO602.1	12.1	Create a schema for unstructured data in MongoDB 1. Select any unstructured dataset in JSON format (e.g., reviews, logs, or profiles) 2. Identify common fields and structure 3. Define a suitable document schema for storing the data in MongoDB	A:31/01/2026 B:05/02/2026 C:30/01/2026			
13	CO602.1	13.1	Run basic aggregation queries on Unstructured dataset in MongoDB 1. Import the dataset created in the previous experiment into a MongoDB collection 2. Run basic aggregation queries to validate the schema (such as counting records, grouping by a field, or calculating averages)	A:07/02/2026 B:12/02/2026 C:06/02/2026			
14	CO602.1	15.1	Install and Configure Hive 1. Install Hive on Hadoop environment 2. Load a dataset into Hive 3. Run basic HiveQL queries (select, insert, delete)	A:11/02/2026 B:16/02/2026 C:10/02/2026			
15	CO602.1	16.1	Execute Data Queries Using HiveQL 1. Load any dataset into Hive 2. Execute joins, group by, and aggregate queries	A:14/02/2026 B:23/02/2026 C:13/02/2026			
16	CO602.1	17.1	Run Pig Scripts for data transformation 1. Load any dataset into Pig 2. Apply filtering and grouping operations	A:18/02/2026 B:26/02/2026 C:17/02/2026			
17	CO602.1	18.1	Execute User Defined Functions (UDF) in Pig for Data Normalization 1. Load any numerical dataset into Pig	A:21/02/2026 B:02/03/2026 C:20/02/2026			

			2. Develop a custom UDF to normalize numerical readings 3. Register the UDF Script Apply the UDF for normalization				
18	CO602.1	19.1	Install Spark on Cluster environment and verify installation with any dataset 1. Install Spark on the cluster 2. Start spark shell 3. Load a sample dataset 4. Verify basic operations	A:25/02/2026 B:05/03/2026 C:24/02/2026			
19	CO602.1	20.1	Perform Data Transformations with RDDs and apply map, filter, reduce operations 1. Load any dataset into an RDD 2. Apply map, filter and reduce operations	A:28/02/2026 B:09/03/2026 C:27/02/2026			
20	CO602.1	21.1	Perform ETL Process on Big Data Using Spark 1. Load dataset from local storage or HDFS 2. Apply basic transformations (such as filtering and aggregation) on the dataset 3. Store the transformed data back into local storage, or HDFS	A:04/03/2026 B:12/03/2026 C:06/03/2026			
21	CO602.1	22.1	Load Structured Dataset and Create Temporary Views in Spark SQL 1. Select and load any structured dataset (such as CSV or JSON) into Spark 2. Inspect the dataset to understand its structure (columns and data types) 3. Create a temporary view from the loaded dataset 4. Display basic records from the view to verify successful creation	A:07/03/2026 B:16/03/2026 C:02/01/2026			
22	CO602.1	26.1	Apply Simple Regression on Large Dataset in Spark	A:11/03/2026 B:16/03/2026			

			1. Load any numerical dataset 2. Build and train a simple regression model 3. Train and apply the model 4. Predict target values	C:10/03/2026			
23	CO602.1	27.1	Apply K-Means Clustering on any Dataset using Spark MLlib 1. Select any structured dataset suitable for clustering (e.g., customer data, product features) 2. Load the dataset into Spark and prepare it for clustering (select relevant numerical features) 3. Apply the K-Means clustering algorithm with a defined number of clusters (K) 4. Output the cluster assignments for each data point 5. Save the clustering result for further use	A:14/03/2026 B:23/03/2026 C:17/03/2026			
24	CO602.1	28.1	Visualize K-Means Clustering Results Using Spark MLlib 1. Load the cluster assignment results from the previous experiment 2. Use visualization tools (such as Matplotlib, Seaborn, or any Spark-compatible library) 3. Plot the clusters on a 2D graph based on two key features 4. Color-code data points according to cluster labels 5. Export the visualization as an image or report	A:25/03/2026 B:30/03/2026 C:27/03/2026			

- **Formative Assessment Criteria:**

Performance Indicators		Weightage
Process Related (15 Marks)		**60%**
1	Logic formation	30%
2	Debugging ability	20%
3	Follow ethical practices	10%
Product Related (10 Marks)		**40%**
1	Expected output	15%
2	Timely Submission	15%
3	Answer to sample questions	10%
	Total (25 Marks)	**100%**

- **Rules for Formative Assessment:**
  1. Formative assessment of each practical is based on Process related (15 marks) and Product related (10 marks) - Total out of 25 marks as per the assessment scheme prescribed in manual given by MSBTE,
  2. Final FormativeAssessment (F.A.) of 25 marks is calculated as follows:
     FA Marks = ((Total obtained marks)*25)/ (25*Total Number of practicals)
  3. Practical Examination will be displayed on Notice board prior to examination.

- **Summative Assessment:**
  There is no summative assessment for laboratory learning

- **Practical wise LLO-CO Mapping:**

Practical No.	LLO	CO301.1	CO301.2	CO301.3	CO301.4	CO301.5
1	1.1	✔				
2	2.1		✔			
3	3.1		✔			
4	4.1		✔			
5	5.1		✔			
6	6.1		✔			
7	7.1		✔			
8	8.1		✔			
9	9.1		✔			
10	10.1			✔		
11	11.1			✔		
12	12.1			✔		
13	13.1			✔		
14	15.1				✔	
15	16.1				✔	
16	17.1				✔	
17	18.1				✔	
18	19.1					✔
19	20.1					✔
20	21.1					✔
21	22.1					✔
22	26.1					✔
23	27.1					✔
24	28.1					✔

- **References:**

Books:

Sr.No	Author	Title	Publisher with ISBN Number
1	Raj Kamal, Preeti Saxena	Big Data Analytics: Introduction to Hadoop, Spark, and Machine-Learning	McGraw Hill Education, New Delhi. ISBN: 9789353164962
2	Seema Acharya, Subhashini Chellappan	Big Data and Analytics	Wiley India Pvt. Ltd., ISBN: 9788126554782
3	M. Vijayalakshmi, Radha Shankarmani	Big Data Analytics	Publication details: Wiley c2017, 2022 N. Delhi Edition: 2nd ed. c2017,ISBN: 9788126565757
4	Holden Karau, Andy Konwinski, Patrick Wendell, Matei Zaharia	Learning Spark: Lightning-Fast Data Analytics	O'Reilly Media Publication Date: January 28, 2015 ISBN-10: 1449358624 ISBN-13: 978-1449358624
5	Pramod J. Sadalage, Martin Fowler	NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence	Addison-Wesley August 10, 2012 ISBN: 978-0321826626
6	Tom White	Hadoop: The Definitive Guide	4th Edition, Released April 2015, Publisher(s): O'Reilly Media, Inc. ISBN: 9781491901632.

Web sites :

Sr. No	Link / Portal	Description
1	https://hadoop.apache.org/	Official website for Apache Hadoop, including documentation, downloads, and tutorials.
2	https://spark.apache.org/	Official website for Apache Spark, providing guides, API references, and use case examples.
3	https://pig.apache.org/	Official site for Apache Pig, with resources for learning Pig Latin and building scripts.
4	https://hive.apache.org/	Official resource for Apache Hive, including installation guides and HiveQL references.
5	https://www.mongodb.com/	MongoDB official site offering documentation, downloads, and free learning courses.
6	https://onlinecourses.nptel.ac.in/noc20_cs92/preview	This course provides an in-depth understanding of terminologies and the core concepts behind big data problems, applications, systems and the techniques that underlie today's big data computing technologies.
7	https://www.tutorialspoint.com/hadoop/index.htm	This brief tutorial provides a quick introduction to Big Data, MapReduce algorithm, and Hadoop Distributed File System.
8	https://www.w3schools.com/mongodb/	This brief tutorial provides a quick introduction to MongoDB.

Tools : Apache Hadoop, MongoDB, Apache Hive, Apache Pig, Apache Spark,

Mrs. J. S. Mahajan
(Name & signature of Staff)

Mrs. R. Y. Thombare
(Name & signature of HOD)

CC: 1. Lab File 2. Course File-BLP 3. Notice Board-AN Lab-01 4. Formative Assessment