DHANALAKSHMI SRINIVASAN ENGINEERING COLLEGE
(AUTONOMOUS)
(Approved by AICTE & Affiliated to Anna University, Chennai)
Accredited with 'A' Grade by NAAC, Accredited by TCS
Accredited by NBA with BME, ECE & EEE
PERAMBALUR - 621 212. Tamil Nadu.
website : www.dsengg.ac.in

LABORATORY COURSEPLAN (2025 – 2026 Even Semester)

LABCOURSE TITLE	BIG DATA ANALYTICS LABORATORY			
LABCOURSECODE	U22AIP52			
LAB COURSE STRUCTURE	LECTURE	TUTORIAL	PRACTICAL	CREDIT
	0	0	4	2
REGULATION	BRANCH	YEAR &SECTION	SEMESTER	ACADEMIC YEAR
2020	IT	III&A, B, C	VI	2025-2026
COURSEINCHARGE				

SYLLABUS

COURSE OBJECTIVE:

- To realize storage of big data using HBase, MongoDB
- To analyze big data using linear models
- To analyze big data using machine learning techniques such as SVM/Decision tree classification and clustering
- To perform visualization of data using plotting frameworks

LISTOFEXPERIMENTS

1. Install, configure, and run Python, NumPy, and Pandas.
2. Install, configure, and run Hadoop and HDFS.
3. Visualize data using basic plotting techniques in Python
4. Implement NoSQL Database Operations: CRUD operations, Arrays using MongoDB.
5. Implement Functions: Count – Sort – Limit – Skip – Aggregate using MongoDB.
6. Implement word count/frequency programs using MapReduce
7. Implement a MapReduce program that processes a dataset.
8. Implement clustering techniques using SPARK
9. Implement an application that stores big data in MongoDB / Pig using Hadoop / R.

TOTAL: 60PERIODS

BIBLIOGRAPHY

TEXT/REFERENCE BOOKS:

- Tom White, 'Hadoop: The Definitive Guide', 4th Edition, O'Reilly, 2015.
- Seema Acharya, Subhashini Chellappan, 'Big Data Analytics', Wiley, 2015.
- Jure Leskovec, Anand Rajaraman, Jeffrey Ullman, 'Mining of Massive Datasets', Cambridge University Press, 2020.
- Ethem Alpaydin, 'Introduction to Machine Learning', 4th Edition, MIT Press, 2020.

HARDWARE:

Standalone desktop / Laptop with at least 8GB RAM

SOFTWARE:

Hadoop, R, Python (scikit-learn, matplotlib, seaborn), MongoDB, HBase

weblinkforresource&Virtuallabreferencelink

https://vsit.edu.in/vlab.html
https://www.balajia.co.in/SBJ/dbmscs8481

EXP. NO.	NAMEOFTHEEXPERIMENTS	NO.OFPERIODS	CUMULATIVEPERIODS
1	Install, configure, and run Python, NumPy, and Pandas.	6	6
2	Install, configure, and run Hadoop and HDFS.	6	12
3	Visualize data using basic plotting techniques in Python	6	18
4	Implement NoSQL Database Operations: CRUD operations and arrays using MongoDB.	6	24
5	Implement Functions: Count – Sort – Limit – Skip – Aggregate using MongoDB.	6	30

6	Implement word count/frequency programs using MapReduce	6	36
7	Implement a MapReduce program that processes a dataset	6	42
8	Implement clustering techniques using SPARK	6	48
9	Implement an application that stores big data in MongoDB / Pig using Hadoop / R.	12	60

COURSE OUTCOME:

Upon completion of the course, the students will be able to:

CO1: Process big data using Hadoop framework (K3)

CO2: Build and apply linear and logistic regression models (K3)

CO3: Perform data analysis with machine learning methods (K3)

CO4: Perform graphical data analysis (K3)

CO5: Store and retrieve big data using NoSQL databases (K3)

CO6: Design and implement mini-projects with IoT and data analytics (K3)

CO-POMapping:

CO	PO 1	PO 2	PO 3	PO 4	PO 5	PO 6	PO 7	PO 8	PO 9	PO 10	PO 11	PO 12
CO1	3	2	1	1	-	-	-	-	-	-	-	-
CO2	3	2	1	1	-	-	-	-	-	-	-	-
CO3	3	2	1	1	-	-	-	-	-	-	-	-
CO4	3	2	1	1	-	-	-	-	-	-	-	-
CO5	3	2	1	1	-	-	-	-	-	-	-	-
CO6	1	3	2	2	-	-	-	-	-	-	-	-
AVG:	**2.67**	**2.17**	**1.17**	**1.17**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**

MODELLABDETAILS:

BATCH	REGISTERNO.	MODE OF LAB CONDUCT	DATE	TIMING
1				

SAMPLE UNIVERSITY QUESTIONS:

1. Install and configure Hadoop in pseudo-distributed mode. Show HDFS directory structure and demonstrate file upload/download operations.
2. Write a MapReduce program in Java/Python to perform **word count** on a given dataset.
3. Implement a MapReduce job to analyze a **weather dataset** and display the maximum temperature per year.
   Build and evaluate a **Linear Regression** model on a sample dataset (e.g., housing prices).
4. Implement **Logistic Regression** for binary classification (e.g., spam email detection).
5. Apply **Decision Tree** classification on a dataset and explain the split criteria (Gini Index / Information Gain).
6. Perform **SVM classification** on an image/text dataset and compare its accuracy with the Decision Tree.
7. Implement **K-Means clustering** on a dataset and visualize the clusters.
8. Store a large dataset in **MongoDB** or **HBase** and retrieve data using queries.
9. Develop a **mini IoT-based data analytics project** (e.g., Smart Weather Monitoring System) and visualize sensor data using Python plotting libraries.

VIVAQUESTIONS:

1. What is Big Data? Explain the 5 V's.
2. What is Hadoop? List its main components.
3. Differentiate between HDFS and MapReduce.
4. What is the role of the Name Node and Data Node in HDFS?
5. Explain the phases of a MapReduce job (Map, Shuffle & Sort, Reduce).
6. Why is MapReduce suitable for large datasets?
7. What are combiners in MapReduce?
8. What is the difference between supervised and unsupervised learning?
9. Compare Linear Regression and Logistic Regression.
10. What is overfitting? How can it be reduced?
11. Explain the working principle of SVM.
12. What is the difference between classification and clustering?
13. What is NoSQL? How is it different from RDBMS?
14. Compare MongoDB and HBase.
15. What are the advantages of using NoSQL in big data?
16. Why is data visualization important in analytics?
17. What are common Python libraries for visualization?

18. Give an example of an IoT application where data analytics is useful.
19. How does data analytics support real-time IoT systems?

Prepared By **VerifiedBy**

ApprovedBy
PRINCIPAL