Program Name : **Diploma in Artificial Intelligence and Machine Learning/ Diploma in Cloud Computing and Big Data**

Program Code : **AN/BD**

Semester : **Sixth**

Course Title : **Big Data Analytics**

Course Code : **22684**

1. RATIONALE

Data analytics techniques enable a business to take raw data and uncover patterns to extract valuable insights. Data analysis helps companies make informed decisions, create a more effective marketing strategy, improve customer experience and streamline operations.

2. COMPETENCY

The aim of this course is to help the student to attain the following ***industry identified*** competency through various teaching learning experiences:

- **Use Big data analytic technologies to process large amount of heterogeneous raw data to retrieve information.**

3. COURSE OUTCOMES (COs)

The theory, practical experiences and relevant soft skills associated with this course are to be taught and implemented, so that the student demonstrates the following industry oriented COs associated with the above mentioned competency:

a. Describe Big data and Big Data Analytics.
b. Apply the Big data Analytics procedure to work on datasets.
c. Describe Hadoop Distributed File System.
d. Analyze structured data using HIVE.
e. Analyze structured, semi structured and unstructured data using SPARK.

4. TEACHING AND EXAMINATION SCHEME

Teaching Scheme (In Hours)				Examination Scheme													
				Theory						Practical							
L	T	P	Credits (L+T+P)	Paper Hrs.	ESE		PA		Total		ESE		PA		Total		
					Max	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max	Min	
3	0	2	5	3	70	28	30*	0	100	40	25@	10	25	10	50	20	

(**) *marks should be awarded on the basis of internal end semester theory exam of 50 marks based on the specification table given in S. No. 9.*

*(~2): For the **practical only courses,** the PA has two components under practical marks i.e. the assessment of practicals (seen in section 6) has a weightage of 60% (i.e.30 marks) and micro-project assessment (seen in section 12) has a weightage of 40% (i.e.20 marks). This is designed to facilitate attainment of COs holistically, as there is no theory ESE.*

Legends: L-Lecture; **T** – Tutorial/Teacher Guided Theory Practice; **P** -Practical; **C** – Credit, **ESE** -End Semester Examination; **PA** - Progressive Assessment, **'#':** No Theory Examination

5. COURSE MAP (with sample COs, Learning Outcomes i.e. LOs and topics)

This course map illustrates an overview of the flow and linkages of the topics at various levels of outcomes (details in subsequent sections) to be attained by the student by the end of the course, in all domains of learning in terms of the industry/employer identified competency depicted at the centre of this map.

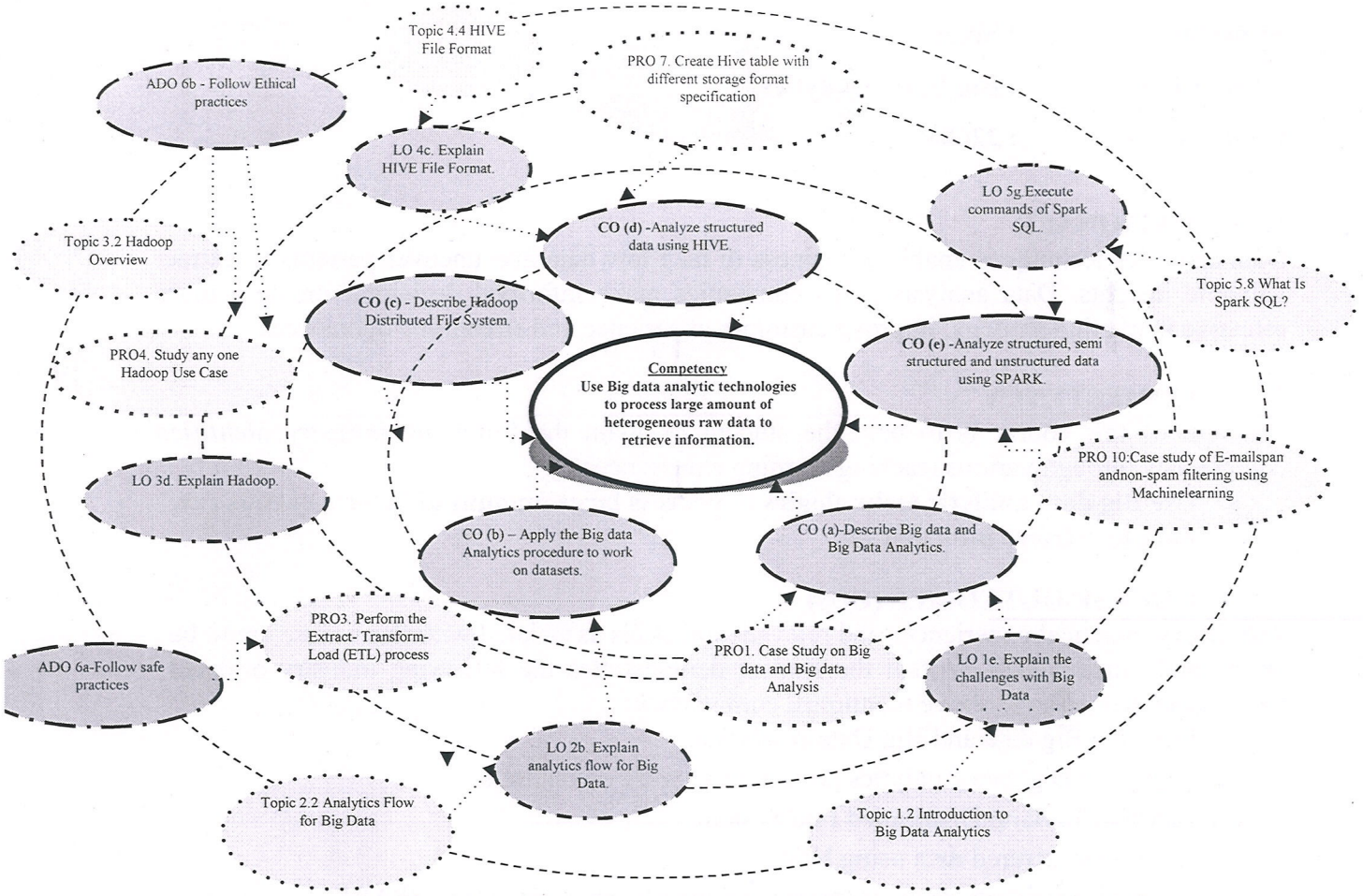

Figure 1 - Course Map

6. SUGGESTED PRACTICALS/ EXERCISES

The practicals/exercises/tutorials in this section are psychomotor domain LOs (i.e.sub-components of the COs) are to be developed and assessed in the student to lead to the attainment of the competency.

Sr. No.	Practical Exercises (Learning Outcomes to be achieved through practical)	Unit No.	Approx. Hrs. Required
1.	Case Study on Big data and Big data Analysis. (Walmart, Uber Netflix, eBay etc.)	I	02*

Sr. No.	Practical Exercises (Learning Outcomes to be achieved through practical)	Unit No.	Approx. Hrs. Required
2.	Write a Pandas program a. To import given excel data into a Pandas Dataframe. b. To get the data types of the given excel data fields. c. To read specific columns from a given excel file. d. To find the sum, mean, max, min value of a specific column of a given excel file. e. To import some excel data skipping some rows or columns. f. To select the specified columns and rows from a given data frame. g. To Delete Rows and Columns from DataFrame.	II	04*
3.	Perform the Extract- Transform-Load (ETL) process a. Import the functions and required modules b. Download the source file c. Extract the zip file d. Set the path for the target files e. Use the extract( ) function to extract data from multiple sources f. Transform the data as per the given requirement using transform( ) function g. Load the data into the target file h. Call the log function for each phase	II	04*
4.	Study any one Hadoop Use Case.	IV	02*
5.	Create Hive table: a. Create Hive External Table. b. Load data into Hive table. c. Create Hive Internal Table.	V	02*
6.	Load the data into Hive Table: a. Load data from Local file system b. Load data from Hdfs file system c. Copy data to Hive table Location d. Sqoop Hive import to import table data	V	04*
7.	Create Hive table with following storage format specification: a. Hive Text File Format b. Hive Sequence File Format c. Hive RC File Format d. Hive AVRO File Format e. Hive ORC File Format f. Hive Parquet File Format	V	02*
8.	Consider the sample logs.txt shown in figure. Write a	VI	04*

Sr. No.	Practical Exercises (Learning Outcomes to be achieved through practical)	Unit No.	Approx. Hrs. Required
	Spark application to count the total number of WARN lines in the logs.txt file. (Implement using Scala / Python Programming) *Sample logs.txt* WARN This is a warning message ERROR This is an error message WARN This is a warning message ERROR This is an error message ERROR This is an error message WARN This is a warning message WARN This is a warning message		
9.	Implement using Scala / Python Programming: a. Create the following data as logdata.log with comma delimiters as shown. 10:24:25,10.192.123.23,http://www.google.com/searchString,ODC1 10:24:21,10.123.103.23,http://www.amazon.com,ODC1 10:24:21,10.112.123.23,http://www.amazon.com/Electronics,ODC1 10:24:21,10.124.123.24,http://www.amazon.com/Electronics/storagedevices,ODC1 10:24:22,10.122.123.23,http://www.gmail.com,ODC2 10:24:23,10.122.143.21,http://www.flipkart.com,ODC2 10:24:21,10.124.123.23,http://www.flipkart.com/offers,ODC1 The schema for these data is Time, IP Address, URL and Location b. Create a DataFrame of the created log file using spark.read.csv.	VI	02*
10.	Write and run SparkSQL queries programmatically for the following requirements. (Implement using Scala / Python Programming) a. How many people accessed the Flipkart domain in each location? b. Who accessed the Flipkart domain in each location? c. List their IpAddress. d. How many distinct Internet users are available in each location? e. List the unique locations available	VI	04*
11.	Read and Write data stored in Apache Hive through Spark SQL. (Implement using Scala / Python Programming)	VI	02*
	Total		**32**

'': compulsory practicals to be performed.*

Note

i. *Given in above tables is suggestive list of practical exercises. Teachers can design other similar exercises.*

ii. *Assessment of the 'Process' and 'Product' related skills in the laboratory/workshop/field work should be done as per suggested sample below:*

Sr. No.	Performance Indicators	Weightage in %
1	Import packages and Libraries of Python / Scala /Hive / Spark.	20
2	Use Python / Scala /Hive / Spark to create, edit, assemble and link the programs.	40
3	Debug, test and execute the programs	20
4	Able to answer oral questions.	10
5	Submission of report in time.	10
	Total	**100**

Additionally, the following affective domain LOs (social skills/attitudes), are also important constituents of the competency which can be best developed through the above mentioned laboratory/field based experiences:

a. Work with various libraries to handle data.
b. Demonstrate working as a leader/a team member.
c. Maintain tools and equipment.
d. Follow ethical practices.

The development of the attitude related LOs of Krathwohl's 'Affective Domain Taxonomy', the achievement level may reach:

- 'Valuing Level' in 1st year
- 'Organizing Level' in 2nd year and
- 'Characterizing Level' in 3rd year.

7. MAJOR EQUIPMENT/ INSTRUMENTS REQUIRED

The major equipment with broad specification mentioned here will usher in uniformity in conduct of experiments, as well as aid to procure equipment by authorities concerned.

S. No.	Equipment Name with Broad Specifications	Expt. S.No.
1.	Hardware: Personal computer, (i3 preferable), RAM minimum 4 GB onwards.	For all Experiments
2.	Operating system: Windows 10 onward	
3.	Software: Editor: Python setup	
	Apache Hadoop and Hive	Practical 5 to 11
4.	Software: Editor: Scala setup	Practical 8 to 11

8. UNDERPINNING THEORY COMPONENTS

The following topics/subtopics should be taught and assessed in order to develop LOs in the cognitive domain for achieving the COs to attain the identified competency.

Unit	Major Learning Outcomes (in cognitive domain)	Topics and Sub-topics
Unit – I Introduction to Big Data Analytics	1a. Describe the characteristics of data. 1b. Define Big Data. 1c. Explain the challenges with Big Data. 1d. Define Big Data Analytics. 1e. Explain the challenges with Big Data Analytics.	1.1 Introduction: • Characteristics of Data • Evolution of big data • Definition of Big Data • Challenges with Big Data • What is Big Data • Why Big Data 1.2 Introduction to Big Data Analytics:

Unit	Major Learning Outcomes (in cognitive domain)	Topics and Sub-topics
	1f. Explain Data Science. 1g. Write down responsibilities of a Data Scientist. 1h. Explain Terminologies Used in Big Data Environment.	• What is Big Data Analytics • Classification of Analytics • Why is Big Data Analytics Important • Data Science • Responsibilities of a Data Scientist • Terminologies Used in Big Data Environments
Unit-II Data Analytics Process	2a. Explain any one Domain specific example of Big Data. 2b. Explain analytics flow for Big Data. 2c. State different Big Data Stack. 2d. Describe mapping analytics flow to Big Data Stack. 2e. State different analytics patterns.	2.1 Domain Specific Examples of Big Data • Web • Financial • Healthcare • Internet of Things • Environment • Logistics & Transportation • Industry • Retail 2.2 Analytics Flow for Big Data • Data Collection • Data Preparation • Analysis Types • Analysis Modes • Visualizations 2.3 Big Data Stack • Raw Data Sources • Data Access Connectors • Data Storage • Batch Analytics • Real-time Analytics • Interactive Querying • Serving Databases, Web &Visualization Frameworks 2.4 Mapping Analytics Flow to Big Data Stack 2.5 Case Study: Genome Data Analysis 2.6 Case Study: Weather Data Analysis 2.7 Analytics Patterns
Unit-III The Big Data Technology: Hadoop	3a. State the features of Hadoop. 3b. Enlist key advantages of Hadoop. 3c. Compare RDBMS versus Hadoop. 3d. Explain Hadoop. 3e. Describe HDFS.	3.1 Introduction to Hadoop: • Features of Hadoop • Key Advantages of Hadoop • Why Hadoop • RDBMS versus Hadoop 3.2 Hadoop Overview 3.3 Use Case of Hadoop 3.4 HDFS 3.5 Processing Data with Hadoop
Unit-IV Introduction to HIVE	4a. State the use of HIVE. 4b. Describe HIVE Architecture. 4c. Explain HIVE File Format. 4d. Execute HIVE Query	4.1 What is HIVE? 4.2 HIVE Architecture 4.3. HIVE Data Types 4.4 HIVE File Format

Unit	Major Learning Outcomes (in cognitive domain)	Topics and Sub-topics
	Language commands. 4e. Explain SERDE. 4f. Describe User Defined Functions.	4.5 HIVE Query Language 4.6 RCFile Implementation 4.7 SERDE 4.8 User Defined Functions
Unit-V Introduction to SPARK	5a. State the use of Apache Spark. 5b. Compare Spark and Hadoop MapReduce. 5c. Describe Apache Spark Architecture. 5d. State the Spark Components. 5e. Define RDD. 5f. tate the RDD Operations. 5g. Execute commands of Spark SQL. 5h. Describe DataFrame Operations. 5i. Describe Generic Load and Save Functions. 5j. Write a code for Building Spark SQL Application with SBT. 5k. Explain Spark Real-Time Use Case.	5.1 What Is Apache Spark? 5.2 Why Apache Spark? 5.3 Spark vs. Hadoop MapReduce 5.4 Apache Spark Architecture 5.5 Spark Components 5.6 Spark Shell 5.7 Spark Core: RDD • RDD Operations • Creating an RDD 5.8 What Is Spark SQL? 5.9 Spark Session 5.10 Creating DataFrames • DataFrame Operations • Dataset Operations 5.11 Different Data Sources: Generic Load and Save Functions 5.12 Building Spark SQL Application with SBT 5.13 Spark Real-Time Use Case • Data Analytics Project Architecture • Use Cases

***Note**: To attain the COs and competency, above listed Learning Outcomes (LOs) need to be undertaken to achieve the 'Application Level' of Bloom's 'Cognitive Domain Taxonomy'.*

9. SUGGESTED SPECIFICATION TABLE FOR QUESTION PAPER DESIGN

Unit No.	Unit Title	Teaching Hours	Distribution of Theory Marks			
			R Level	U Level	A Level	Total Marks
I	Introduction to Big Data Analytics	08	08	04	-	12
II	Data Analytics Process	10	06	06	02	14
III	The Big Data Technology: Hadoop	08	06	06	02	14
IV	Introduction to HIVE	10	02	06	06	14
V	Introduction to SPARK	12	06	04	06	16
	Total	**48**	**28**	**26**	**16**	**70**

Legends: R=Remember, U=Understand, A=Apply and above (Bloom's Revised taxonomy)
Note: This specification table provides general guidelines to assist student for their learning and to teachers to teach and assess students with respect to attainment of LOs. The actual distribution of marks at different taxonomy levels (of R, U and A) in the question paper may vary from above table.
This specification table also provides a general guideline for teachers to frame internal end semester practical theory exam paper which students have to undertake.

10. SUGGESTED STUDENT ACTIVITIES

Other than the classroom and laboratory learning, following are the suggested student-related ***co-curricular*** activities which can be undertaken to accelerate the attainment of the various outcomes in this course:

a. Prepare journals based on practical performed in laboratory.
b. Library/E-Book survey regarding assembly language programming used in Computer industries.
c. Prepare power point presentation for showing different types of Assembly language Programming Applications.

11. SUGGESTED SPECIAL INSTRUCTIONAL STRATEGIES (if any)

These are sample strategies, which the teacher can use to accelerate the attainment of the various outcomes in this course:

a. Massive open online courses (***MOOCs***) may be used to teach various topics/sub topics.
b. ***'L' in item No. 4*** does not mean only the traditional lecture method, but different types of teaching methods and media that are to be employed to develop the outcomes.
c. About ***15-20% of the topics/sub-topics*** which is relatively simpler or descriptive in nature is to be given to the students for ***self-directed learning*** and assess the development of the LOs/COs through classroom presentations (see implementation guideline for details).
d. With respect to item No.10, teachers need to ensure to create opportunities and provisions for ***co-curricular activities***.
c. Guide student(s) in undertaking micro-projects.
f. No. of practical's selection to be performed should cover all units.

12. SUGGESTED MICRO-PROJECTS

Only one micro-project is planned to be undertaken by a student assigned to him/her in the beginning of the semester. S/he ought to submit it by the end of the semester to develop the industry oriented COs. Each micro-project should encompass two or more COs which are in fact, an integration of practicals, cognitive domain and affective domain LOs. The micro-project could be industry application based, internet-based, workshop-based, laboratory-based or field-based. Each student will have to maintain a dated work diary consisting of individual contributions in the project work and give a seminar presentation of it before submission. The total duration of the micro-project should not be less than ***16 (sixteen) student engagement hours*** during the course.

In the first four semesters, the micro-project could be group-based. However, in higher semesters, it should be individually undertaken to build up the skill and confidence in every student to become problem solver so that s/he contributes to the projects of the industry. A suggestive list is given here. Similar micro-projects could be added by the concerned faculty:

a. Study of Hadoop in the Financial Sector/ Healthcare Sector/ Retail Sector/for Telecom Industry/for Building Recommendation System.

b. Load the data set and store it in a data-frame using Pandas and perform following operations.

- Remove the missing data using List-wise deletionRemove the missing data using Pair-wise deletion
- Remove the missing data using Forward filling
- Check for duplicate value
- Separate categorical and numerical data.
- To work with Missing values using functions {isnull(), notnull(), dropna(), fillna(), replace(), interpolate()}.

c. Write a Pandas program

- To join the two given dataframes along columns.

- To join the two given dataframes along rows and merge with another dataframe along the common column.
- To join the two dataframes using the common column of both dataframes.

d. Create Hive table, load data into Hive table and Execute following Hive built-in functions
on given Hive Table (Simple Functions, Aggregate Functions, Date Function).

e. Create an RDD:

- Use the parallelize method of SparkContext. Create Array of integers and pass that as an argument to the parallelize method.
- Using an external data source.
- Using an external datasource HDFS
- Create an RDD of a numeric list. Then apply map(func) to multiply each element by 2.

f. Implement Matrix algorithms in SparkSql programming.

g. Perform Untyped Dataframe operations of SparkSQL (Select, Filter and Aggregate Operations) on a given dataset.

13. SUGGESTED LEARNING RESOURCES

S. No.	Title of Book	Author	Publication
1	Big Data And Analytics Second Edition	Seema Acharya Subhashini Chellappan	Wiley India ISBN: 978-81-265-7951-8 ISBN: 978-81-265-8836-7(ebk)
2	Big Data Science & Analytics A Hands-On Approach	Arshdeep Bahga Vijay Madisetti	ISBN: 978-1-949978-00-1
3	Data Analytics Using Python First Edition	Bharti Motwani	Wiley India ISBN: 978-81-265-0295-0 ISBN: 978-81-265-8965-4(ebk)
4	Practical Apache Spark - Using the Scala API	Subhashini Chellappan Dharanitharan Ganesan	Apress ISBN-13 (pbk): 978-1-4842-3651-2 ISBN-13 (electronic): 978-1-4842-3652-9

14. SOFTWARE/LEARNING WEBSITES

a. https://spark.apache.org/docs/latest/rdd-programmingguide.html (For practicals on
b. Spark) (As on 18 April 2023)
c. https://www.simplilearn.com/what-is-big-data-analytics-article (As on 18 April 2023)
d. https://www.analyticsvidhya.com/blog/2021/06/implementing-python-to-learn-data-engineering-etl-process/ (As on 18 April 2023)