

**(Established under the Presidency University Act, 2013 of the Karnataka Act 41 of 2013)**

**ACA-2[2022] COURSE HAND OUT [Integrated Course]**

**SCHOOL: SOE DATE OF ISSUE: 23rd March, 2022**

**DEPT. : CSE**

**NAME OF THE PROGRAM : B. Tech**

**P.R.C. APPROVAL REF. : PU/AC-17/IST/2019-2023/2021**

**SEMESTER/YEAR : VI/ III Year**

**COURSE CODE & TITLE : CSE219 - Big Data Analytics**

**COURSE CREDIT STRUCTURE (L-P-C): 1-4-3**

**CONTACT HOURS : 45 hours**

**COURSE INSTRUCTOR :** Dr.S.Senthilkumar, Mr Amogh Kulkarni, Ms Pavithra.N,

Dr.S.P.Anandaraj, Ms.Ila Chandrakar, Ms Aemi Kalaria,

Mr Raghavendra.M.Devadas, Ms Bhumika.

**PROGRAM OUTCOMES :**

Graduates of the B. Tech. Program in Computer Science and Engineering will be able to:

**PO 1: Engineering knowledge: Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems. [M]**

**PO 2: Problem analysis: Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences. [H]**

**PO-3: Design/development of solutions: Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations. [H]**

P0-4: Conduct investigations of complex problems**:** Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.

**PO-5 Modern tool usage: Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations [H]**

PO-6 The engineer and society: Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to the professional engineering practice.

PO-7 Environment and sustainability**:** Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development.

PO-8 Ethics: Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.

**PO-9 Individual and team work: Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings. [M]**

**PO-10 Communication: Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions. [L]**

**PO-11 Project management and finance: Demonstrate knowledge and understanding of the engineering and management principles and apply these to one’s own work, as a member and leader in a team, to manage projects and in multidisciplinary environments. [L]**

PO-12 Life-long learning: Recognize the need for, and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change.

**COURSE PREREQUISITES:**

DDL, DML of SQL Queries and Creation of Class & object, interface, reading & writing a file, control statements in java programming.

**COURSE DESCRIPTION:**

This course is designed to provide the fundamental knowledge to equip students being able to handle real world big data problems including the three key resources of Big Data: people, organizations and sensor. With the advancement of IT storage, processing, computation and sensing technologies, Big data has become a novel norm of life.

Topics include: Big data and its importance, Four v’s, big data analytics and applications. Big data technologies, data discovery, open source technology for big data analytics, cloud and big data, predictive analytics, mobile business intelligence and big data ,crowd sourcing analytics, inter- and trans-firewall analytics, information management, Processing big data, Hadoop map reduce, implementation of big data analytics, big data convergence, analytics business maturity model. Big data tools and techniques, installing and user defined functions, data processing operators, Sample projects using tools.

**Type of Skill:** Employability /Entrepreneurship

**Nature of the Course:** Application Based

**COURSE OUTCOMES:**

This course is aimed at assisting students to comprehend the concepts of big data to examine large and varied data sets through open source tools using the Hadoop framework.

**After the completion of the course students shall be able to:**

CO1: Describe the fundamental concepts of big data analytics (Knowledge)

CO2: Apply Map-Reduce programming on the given datasets to extract required insights. (Application).

CO3: Employ appropriate Hadoop Ecosystem tools such as Hive, Hbase to perform data analytics for a given problem (Application)

CO4: Use Spark tool to analyze the given dataset for a given problem. (Application).

**MAPPING OF C.O. WITH P.O.**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **CO No.** | **PO-1** | **PO-2** | **PO-3** | **PO-4** | **PO-5** | **PO-6** | **PO-7** | **PO-8** | **PO-9** | **PO-10** | **P0-11** | **P0-12** |
| **1** | **H** | **-** | **-** | **-** | **-** | **-** | **-** | **-** | **-** | **L** | **-** | **-** |
| **2** | **M** | **H** | **H** | **-** | **H** | **-** | **-** | **-** | **M** | **L** | **L** | **-** |
| **3** | **M** | **H** | **H** | **-** | **H** | **-** | **-** | **-** | **M** | **L** | **L** | **-** |
| **4** | **M** | **H** | **H** | **-** | **H** | **-** | **-** | **-** | **M** | **L** | **L** | **-** |

**COURSE CONTENT (SYLLABUS):**

**Module 1:Introduction to Big data Analytics [Knowledge] ( 10 sessions [P-8; L-2])**

Introduction to Big Data:Basics of Distributed File System, Four Vs, Drivers for Big data, Big data applications, Structured, unstructured, semi-structured and quasi structured data. Big data Challenges-Traditional versus big data approach, The Big Data Technology Landscape: No-SQL

The Hadoop: History of Hadoop-Hadoop use cases, The Design of HDFS, Blocks and replication management, Rack awareness, HDFS architecture, HDFS Federation, Name node and data node, Anatomy of File write, Anatomy of File read.

**Module 2:Hadoop MapReduce Framework [Application] (10 sessions [P-8; L-2])**

**Mapreduce:** Hadoop Map Reduce paradigm, Map and reduce tasks, Job Tracker and task tracker, Map reduce execution pipeline, Key value pair, Shuffle and sort, Combiner and Partitioner, APIs used to Write/Read files into/from Hadoop. Hadoop 2.0 Features, Name Node High Availability, YARN Architecture.

**Module 3: Hive and Hbase Analytical tools [Application] (10 sessions [P-8; L-2])**

**Hive :** Apache Hive with Hive Installation, Hive Data Types, Hive Table partitioning, Hive DDL commands, Hive DML commands, and Hive sort by vs. order by, Hive Joining tables, Hive bucketing.

**Hbase** : Introduction to HBase and its working architecture- Commands for creation and listing of tables- disabled and is disabled of table - enable and is enabled of table- describing and dropping of table-Put and Get command - delete and delete all command-commands for scan, count, truncate of tables.

**Module 4: Data Analytics with Spark [Application] (10 sessions [P-8; L-2])**

**Spark:** Introduction to Apache Spark A unified Spark, Who uses Spark and for what?, A Brief History of Spark, Spark version and releases, Storage layers for Spark. Programming with RDDs: RDD Basics, Creating RDDs, RDD Operations, Passing functions to Spark, Common Transformations and Actions, Persistence. Scala: The Basics, Control Structures and functions, Working with arrays, Maps and Tuples.

**DELIVERY PROCEDURE (PEDAGOGY):**

|  |  |  |
| --- | --- | --- |
| **METHOD** | **TOPIC** | **REMARKS** |
| Student Centric Methods | Mapreduce- Different datasets are given to solve problem statement. | Dataset resource link:  <https://www.kaggle.com/datasets>  https://data.world/datasets/open-data |
| Self-learning topics | Built-in Functions of hadoop mapreduce framework in java and basics of scala | Book: Big Data, Black Book: Covers Hadoop 2, MapReduce, Hive, YARN, Pig, R and Data Visualization  DT Editorial Services  ISBN: 9789351199311 |
| Online course based learning | NPTEL: Big Data Computing (8 Weeks ) (https://onlinecourses.nptel.ac.in/noc20\_cs92/preview  Coursera :1. Big Data Essentials: HDFS, MapReduce and Spark RDD ( 6 weeks)  2. Introduction to BigData(17 hours): <https://www.coursera.org/learn/big-data-introduction>,  EDX: Big Data Fundamentals (10 hours): <https://www.edx.org/course/big-data-fundamentals> | Students are advised to carryout certification |
| Experiential Learning: | Students should complete lab programs by the end of each practical session and module-wise (1,2) assignments before deadline. | Implement dataset analysis with different bigdata tools. |

**Textbook:**

1. Big Data and Analytics- Seema Acharya, Subhashini Chellappan-2015, 2nd Edition, Wiley Publication.
2. Analytics in a Big data world- Bart Baesens- 2nd Edition, Wiley Publication. 2018

**Reference Books:**

1. Big data Analytics, Radha Shankarmani and vijayalakshmi second edition wiley publication 2015
2. Big Data, Anil Maheshwari , McGraw Hill education 2017
3. Tom White, “Hadoop: The Definitive Guide”, 3rd Edition, O’reilly. 2016

**Online Resources:**

1. https://onlinecourses.nptel.ac.in/noc20\_cs92/preview
2. <https://www.coursera.org/learn/big-data-introduction>,
3. <https://www.edx.org/course/big-data-fundamentals>

**GUIDELINES TO STUDENTS:**

1. Follow the instructions of course instructor in both class and lab.
2. Avoid being absent in labs as it will affect the understanding of the experiments.

**SKILL SETS TO BE DEVLOPED:**

# An attitude of enquiry.

* 1. Confidence and ability to tackle new problems.
  2. Ability to interpret events and results.
  3. Ability to work as a leader and as a member of a team.
  4. Assess errors in systems/processes/programs/computations and eliminate them.
  5. Observe and measure physical phenomena.
  6. Write reports.
  7. Select suitable equipment, instrument, materials & software
  8. Locate faults in system/Processes/software.
  9. Manipulative skills for setting and handling systems/Process/ Issues
  10. The ability to follow standard /Legal procedures.
  11. An awareness of the Professional Ethics.
  12. Need to observe safety/General precautions.
  13. To judge magnitudes/Results/issues without actual measurement/actual contacts

**MAPPING OF SKILL SET WITH P.O**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **PO-1** | **PO-2** | **PO-3** | **PO-4** | **PO-5** | **PO-6** | **PO-7** | **PO-8** | **PO-9** | **PO-10** | **PO-11** | **PO-12** |
| **SK1** | **✓** | **✓** | **✓** | **-** | **-** | **-** | **-** | **-** | **✓** | **✓** | **✓** | **-** |
| **SK2** | **✓** | **✓** | **✓** | **-** | **-** | **-** | **-** | **-** | **✓** | **-** | **✓** | **-** |
| **SK3** | **-** | **✓** | **✓** | **-** | **✓** | **-** | **-** | **-** | **✓** | **✓** | **✓** | **-** |
| **SK4** | **-** | **✓** | **✓** | **-** | **✓** | **-** | **-** | **-** | **✓** | **-** | **✓** | **-** |
| **SK5** | **-** | **-** | **-** | **-** | **-** | **-** | **-** | **-** | **✓** | **-** | **✓** | **-** |
| **SK6** | **-** | **-** | **-** | **-** | **-** | **-** | **-** | **-** | **✓** | **-** | **✓** | **-** |
| **SK7** | **-** | **✓** | **✓** | **-** | **✓** | **-** | **-** | **-** | **✓** | **-** | **✓** | **-** |
| **SK8** | **-** | **-** | **-** | **-** | **-** | **-** | **-** | **-** | **✓** | **-** | **✓** | **-** |
| **SK9** | **-** | **✓** | **✓** | **-** | **-** | **-** | **-** | **-** | **✓** | **-** | **✓** | **-** |
| **SK10** | **-** | **-** | **-** | **-** | **-** | **-** | **-** | **-** | **-** | **-** | **✓** | **-** |

**COURSE CONTENT & TASK SCHEDULE FOR LABORATORY COMPONENT:**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Sl. No.** | **Session Number**  **and Date** | **Task No** | **Task** | **Level 01** | **Level 02** | **Number of Lab Sessions required to complete the task** | **Skills to be developed** | **Course Outcome to be developed** |
| **01** | **P1 25/03/2022** | 1 | Introduction to Hadoop Ecosystem tools | Overview of Hadoop Ecosystem tools | **-** | **2** | SK1, SK2 | CO1 |
| **02** | **P2 31/03/2022** | 2 | Introduction to Hadoop distributed file System. | Basics of HDFS | **-** | **2** | SK1, SK2 | CO1 |
| **03** | **P3 07/04/2022** | 3 | Installation of Hadoop single node cluster using Ubuntu operating system. | Hadoop Single node cluster installation on ubuntu | Hadoop Multiple node cluster installation, windows installation | **3** | SK2, SK3, SK4, SK5, SK6, SK7, SK8, SK9 | CO1 |
| **04** | **P4**  **21/04/2022** | 4 | Working with Hadoop Commands | Basic hadoop commands | Cluster administration/  configuration commands | **1** | SK2, SK3, SK4, SK5, SK6, SK7, SK8, SK9 | CO1 |
| **05** | **P5**  **28/4/2022** | 5 | Introduction to Mapreduce framework | Basics of Mapreduce | - | **1** | SK2, SK3, SK4, SK5, SK6, SK7, SK8, SK9 | CO2 |
| **06** | **P6**  **29/4/2022** | 6 | Word Count analysis using sample data set (MapReduce) | Word count analysis for given dataset | Analysis on particular matching word on huge dataset | **2** | SK2, SK3, SK4, SK5, SK6, SK7, SK8, SK9 | CO2 |
| **07** | **P7 05/05/2022** | 7 | Stock analysis using sample data set (MapReduce) | Web log analysis on server log dataset | Analysis on particular data and time , user , matching IP address with missing values | **2** | SK2, SK3, SK4, SK5, SK6, SK7, SK8, SK9 | CO2 |
| **08** | **P8 06/05/2022** | 8 | Web log analysis using sample data set (MapReduce) | Stock analysis on given dataset | Analysis with max, min, average functions on particular field with missing values | **2** | SK2, SK3, SK4, SK5, SK6, SK7, SK8, SK9 | CO2 |
| **08** | **P9 12/05/2022** | 9 | Temperature analysis using sample data set .(MapReduce) | Temperature analysis on given dataset | Analysis with max, min, average functions on particular field with missing values | **2** | SK2, SK3, SK4, SK5, SK6, SK7, SK8, SK9 | CO2 |
| **09** | P10  **19/05/2022** | 10 | Working on basic hive commands -1. | Basic hive commands for given dataset | External table , static and dynamic partitioning with given dataset | **2** | SK2, SK3, SK4, SK5, SK6, SK7, SK8, SK9 | CO3 |
| **10** | P10 **27/05/2022** | 11 | Working on advance hive commands-2. | Advance hive commands | Bucket creation with given size | **2** | SK2, SK3, SK4, SK5, SK6, SK7, SK8, SK9 | CO3 |
| **11** | P9 **03/06/2022** | 12 | Working on basic hbase commands -1. | Basic Hbase commands | Retrieval and deletion based on timestamp | **2** | SK2, SK3, SK4, SK5, SK6, SK7, SK8, SK9 | CO3 |
| **12** | P10 **09/06/2022** | 13 | Working on advance hbase commands-2. | Advance Hbase commands | Cluster Replication Commands | **2** | SK2, SK3, SK4, SK5, SK6, SK7, SK8, SK9 | CO3 |
| **13** | P11 **16/06/2022** | 14 | Install, Deploy & configure Apache Spark | Installation of spark | Spark installation on windows | **2** | SK2, SK3, SK4, SK5, SK6, SK7, SK8, SK9. SK10 | CO4 |
| **14** | P12  **17/06/2022** | 15 | Word count analysis using RDD and FlatMap | word count analysis | Analysis on different dataset stored in HDFS. | **2** | SK2, SK3, SK4, SK5, SK6, SK7, SK8, SK9, SK10 | **C**O4 |

**COURSE SCHEDULE FOR THEORY COMPONENT:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sl.**  **No.** | **Activity** | **Starting Date** | **Concluding Date** | **Total Number of Periods** |
| **01** | **Over View of the course** | **23/03/2022** | **23/03/2022** | **1** |
| **02** | **Module : 01** | **30/03/2022** | **13/04/2022** | **4** |
| **03** | **Module: 02** | **27/4/2022** | **4/05/2022** | **2** |
| **04** | **Module:03** | **11/05/2022** | **18/05/2022** | **2** |
| **05** | **Module:04** | **01/06/2022** | **10/06/2022** | **3** |

**COURSE SCHEDULE FOR LABORATORY COMPONENT:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sl. No.** | **Activity** | **Starting Date** | **Concluding Date** | **Total Number of Periods** |
| **01** | **Over View of the course** | **23/03/2022** | **23/03/2022** | **1** |
| **02** | **Laboratory Familiarization** | **24/03/2022** | **24/03/2022** | **1** |
| **02** | **Demonstration of first set of Experiments/Skills** | **25/03/2022** | **06/05/2022** | **14** |
|  | **Conduct of first set of experiments** |
| **03** | **Assignment/ Test or any other activity** | **12/05/2022** | **13/05/2022** | **2** |
| **04** | **Conduct of second set of experiments** | **19/05/2022** | **16/06/2022** | **14** |
| **05** | **Summary of the Laboratory tasks** | **17/06/2022** | **17/06/2022** | **2** |
| **06** | **End Term Evaluation** | **As per COE schedule** |  |  |

**SCHEDULE OF INSTRUCTION FOR THE THEORY/LABORATORY COMPONENT:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Sl. No** | **Session**  **no**  **[date if possible]** | **Lesson Title** | **Topics** | **Course Outcome Number** | **Delivery Mode &Tools used** | **Reference** |
| **1** | **L1 23/03/2022** | **MODULE 1: Introduction to Bigdata Analytics** | Introduction to course handout and evaluation of the course, Program integration | **CO1** | **PPTS/ chalk & talk /online** | **T1,R1** |
| **2** | **P1**  **24/03/2022** | Basics of Distributed File System, Four Vs Drivers for Big data, Big data applications, Structured, unstructured, semi-structured and quasi structured data | **PPTS/ chalk & talk /online** | **T1,R1** |
| **3** | **P2**  **25/03/2022** | Introduction to Hadoop Ecosystem tools | **PPTS/ chalk & talk /online** | **T1,R1** |
| **4** | **L2**  **30/03/2022** | Big data Challenges-Traditional versus big data approach | **PPTS/ chalk & talk /online** | **T1,R1** |
| **5** | **P3**  **31/03/2022** | Introduction to Hadoop distributed file System. | **PPTS/ chalk & talk /online** | **T1,R1** |
| **6** | **P4**  **31/03/2022** | Installation of Hadoop single node cluster using Ubuntu operating system. | **PPTS/ chalk & talk /online** | **T1,R1** |
| **7** | **L3**  **06/04/2022** | The Big Data Technology Landscape: No-SQL | **PPTS/ chalk & talk /online** | **T1,R1** |
| **8** | **P5**  **21/04/2022** | Working with Hadoop Commands. | **PPTS/ chalk & talk /online** | **T1,R1** |
| **9** | **P6**  **22/04/2022** |  | History of Hadoop-Hadoop use cases, The Design of HDFS, Blocks and replication management |  | **PPTS/ chalk & talk /online** | **T1,R1** |
| **10** | **L4**  **27/04/2022** | Rack awareness, HDFS architecture, HDFS Federation, Name node and data node, Anatomy of File write, Anatomy of File read.  Course integration | **PPTS/ chalk & talk /online** | **T1,R1** |
| **11** | **P7**  **28/04/2022** | **MODULE 2: Hadoop MapReduce Framework** | Introduction to Mapreduce framework | **CO 2** | **PPTS/ chalk & talk /online** | **T1,R1** |
| **12** | **P8**  **//2022** | **Mapreduce:** Hadoop Map Reduce paradigm, Map and reduce tasks, Job Tracker and task tracker, | **PPTS/ chalk & talk /online** | **T1,R1** |
| **13** | **L5**  **//2022** | Map reduce execution pipeline, Key value pair, Shuffle and sort, Combiner and Partitioner, | **PPTS/ chalk & talk /online** | **T1,R1** |
| **14** | **P9** | Assesment 1 | **PPTS/ chalk & talk /online** | **T1,R1** |
| **15** | **P10**  **29/4/2022** | Word count analysis using sample data set(mapreduce) | **PPTS/ chalk & talk /online** | **T1,R3** |
| **16** | **L6**  **04/05/2022** | APIs used to Write/Read files into/from Hadoop. Hadoop 2.0 Features, Name Node High Availability, YARN Architecture. |  |  |
| **17** | **P11**  **05/05/2022** | Stock analysis using sample data set(mapreduce) | **PPTS/ chalk & talk /online** | **T1,R3** |
| **18** | **P12**  **06/05/2022** | Web log analysis using sample data set (mapreduce) | **PPTS/ chalk & talk /online** | **T1,R3** |
| **19** | **L8**  **11/05/2022** | Revision of previous class | **PPTS/ chalk & talk /online** | **T1,R3** |
| **20** | **P13**  **12/05/2022** | Temperature analysis using sample data set(mapreduce)  Couse integration | **PPTS/ chalk & talk /online** | **T1,R3** |
| **22** | **P14** |  | Conduction of mid-term lab examination |  |  |
| **23** | **L9**  **18/05/2022** | **Module – 3 Hive and Hbase Analytical tools** | Apache Hive with Hive Installation, Hive Data Types, Hive Table partitioning, Hive DDL commands | **CO 3** | **PPTS/ chalk & talk /online** | **T1,R3** |
| **24** | **P15**  **19/05/2022** | Working on basic hive commands -1 | **PPTS/ chalk & talk /online** | **T1,R3** |
| **25** | **P16**  **27/05/2022** | Working on advance hive commands-2 | **PPTS/ chalk & talk /online** | **T1,R3** |
| **27** | **L10**  **01/06/2022** | Hive DML commands, and Hive sort by vs. order by, Hive Joining tables, Hive bucketing. | **PPTS/ chalk & talk /online** | **T1,R3** |
| **28** | **P17**  **02/06/2022** | Assessment-2 | **PPTS/ chalk & talk /online** | **T1,R3** |
| **29** | **P18**  **03/06/2022** | Working on basic hbase commands -1 |  | **PPTS/ chalk & talk /online** | **T1,R3** |
| **30** | **L11**  **08/06/2022** | Introduction to HBase and its working architecture- Commands for creation and listing of tables- disabled and is disabled of table - enable and is enabled of table | **PPTS/ chalk & talk /online** | **T1,R3** |
| **31** | **P19**  **09/06/2022** | Working on advance hbase commands-2 | **PPTS/ chalk & talk /online** | **T1,R3** |
| **32** | **L12**  **10/6/2022** | describing and dropping of table-Put and Get command - delete and delete all command-commands for scan, count, truncate of tables.  Couse integration | **PPTS/ chalk & talk /online** | **T1,R3** |
| **35** | **P20**  **15/06/2022** | **Module 4: Data Analytics with Spark** | Introduction to Apache Spark A unified Spark, Who uses Spark and for what?, A Brief History of Spark, Spark version and releases, Storage layers for Spark | **CO4** | **PPTS/ chalk & talk /online** | **T1,R3** |
| **36** | **P21**  **16/06/2022** | Programming with RDDs: RDD Basics, Creating RDDs, RDD Operations, Passing functions to Spark, Common Transformations and Actions, Persistence | **PPTS/ chalk & talk /online** | **T1,R3** |
| **37** | **L12**  **17/06/2022** | Scala: The Basics, Control Structures and functions, Working with arrays, Maps and Tuples. | **PPTS/ chalk & talk /online** | **T1,R3** |
| **38** | **P22**  **18/06/2022** | Install, Deploy & configure Apache Spark | **PPTS/ chalk & talk /online** | **T1,R3** |
| **39** | **P23**  **19/06/2022** | Word count analysis using RDD and FlatMap  Benefits of the course to students. Program Integration | **PPTS/ chalk & talk /online** | **T1,R3** |
| **40** | **P24**  **20/06/2022** | Assessment -3 | **PPTS/ chalk & talk /online** | **T1,R3** |

**ASSESSMENT SCHEDULE FOR THEORY COMPONENT:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Sl.**  **No.** | **Assessment type**  **[Include here assessment method for self-learning component also]** | **Contents** | **Course outcome**  **Number** | **Duration**  **(In**  **Hours)** | **Marks** | **Weightage** | **Venue, Date & Time** |
| **1** | **Assesment 1:Quiz** | Module -1 | 1 | 20 MIN | 20 | 10% | Lab |
| **2** | **Midterm exam** | Module 1,2 | 1,2 | 2 HOURS | 15 | 5% | Lab According to COE Schedule |
| **3** | **End term exam** | Module 1,2,3,4 | 1,2,3,4 | 3 HOURS | 20 | 10% | Lab According to COE Schedule |

**ASSESSMENT DETAILS FOR LABORATORY COMPONENT:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Sl.**  **No.** | **Assessment type**  **[Include here assessment method for self-learning component also]** | **List of Tasks** | **Course outcome**  **Number** | **Duration**  **In Hours** | **marks** | **weightage** | **Venue, DATE &TIME** |
| **1** | **Assessment -2** | Module -2 | 2 | 2 HOURS | 20 | 10% | Lab |
| **2** | **Assessment-3** | Module -3,4 | 3 | 2 HOURS | 20 | 10% | Lab |
| **3** | **Mini Project** | Module -2,3,4 | 2,3,4 | 2 HOURS | 30 | 15% | Lab |
| **OR** | | | | | | | |
|  | **Case Study/**  **Assignment** | Module 2,3,4 | 2,3,4 | NA | 30 | 15% | NA |
| **4** | **Record evaluation** | Module 1,2,3,4 | 1,2 3,4 | - | 10 | 5% | Lab |
| **5** | **Mid Term Lab Practical** | Module 1,2 | 1,2 | 2 HOURS | 25 | 15% | Lab According to COE Schedule |
| **6** | **End Term Examinations ( Lab and Theory)** | Module 1,2,3,4 | 1,2,3,4 | 3 HOURS | 40 | 20% | Lab According to COE Schedule |

**Assessment Matrix for Daily Task Evaluation for Laboratory component:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Sl. No.** | **Task No.** | **Marks for activity 01 (3)** | **Marks for activity 02 (3)** | **Marks for activity 03 (4)** | **Total Marks (Record)** |
| **1** | **3,4** | **Installation of vmware, ubuntu** | **Set configuration files** | **Installation of hadoop single node cluster** | **10** |
| **2** | **6,7,8,9** | **Creation of mapreduce class** | **Create and store dataset on hdfs** | **Execution of hadoop commands to generate mapreduce output** | **10** |
| **3** | **10,11** | **Execution of basic hive commands** | **Create and store dataset on hive** | **Execution of advance hive commands** | **10** |
| **4** | **12,13** | **Execution of basic hbase commands** | **Create and store dataset on hbase** | **Execution of advance hbase commands** | **10** |
| **5** | **14,15** | **Installation of spark on ubuntu** | **Develop scala program for word count analysis, Flat Map** | **Execution of scala program** | **10** |

**COURSE CLEARANCE & EVALUATION CRITERIA:**

1. Students are advised to maintain minimum attendance of 75% in this course. Failing which, the student will not be permitted to attend the end term examination and subsequently awarded “NP” grade.

2. The student needs to score minimum 40 marks out of 100 (40%) in continuous assessment. Failing which, the student will be awarded “NE” (not eligible) grade to appear for the end term examination.

3. The student needs to score minimum 30 marks out of 100 (30%) in end term assessment. Failing which, the student will be awarded “F” (fail) grade.

The Student must have secured a minimum of 40% (i.e., 80 marks out of 200) of the AGGREGATE of the marks/weightage of the components of the Pre End Term Marks and the End Term Final Examination. Otherwise will be awarded “F” (fail) Grade.

If the student misses an evaluation component, he/she may be granted a make-up. In case of an absence that is foreseen, make-up request should be personally made to the Instructor-in-Charge, well ahead of the scheduled evaluation component. Reasons for unanticipated absence that qualify a student to apply for make-up include medical emergencies or personal exigencies. In such an event, the student should contact the Instructor-in-Charge as soon as practically possible.

**CONTACT TIMINGS IN THE CHAMBER FOR ANY DISCUSSIONS:**

Contact timings for any discussions will be informed in the class.

**Sample Thought Provoking Questions [For Theory Component]:**

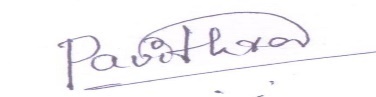
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sl No.** | **Question** | **Marks** | **Course Outcome No.** | **Bloom’s**  **Level** |
| **1** | A client is running his business on Mysql. He want to migrate his data from Mysql platform to HIVE. As a developer what will be your choice of tool. Highlight important features of the tool | 5 | CO1 | L1 |
| 2 | Huge amount of sensor data is generated from different courses which was unstructured in form. They moved to Hadoop framework for storing and analyzing data. What technology in Hadoop framework, they can use to analyse this unstructured data? | 10 | CO2 | L2 |
| 3 | A client wants a database to be developed for his business. He is not sure of how many attributes are required for a particular table. He tells that number of attributes for every row may vary and he is also sure that he may need thousands of attributes for the said table. (i.e it will be a very big table). Which tool best fits for developing the clients requirement. | 10 | CO3 | L2 |
| 4 | Shravani is working as a data scientist in Mu Sigma. She is working with a team of ten members in a project. 500 TB of census data need to be analyzed by the team. An interactive interface has to fetch the required results from the data file. Identify the best suitable analytical tool and explain how the census data can be analyzed. | 10 | CO4 | L2 |

**Sample Thought Provoking Questions to be asked to Assess the Students’ Preparedness to carry out the Task [For Laboratory Component] :**

|  |  |  |  |
| --- | --- | --- | --- |
| **Sl No.** | **Question** | **Task No.** | **Course Outcome No.** |
| 1 | The Analyst wants to know what is the average number of views for each YouTube channel so that trending videos and channel can be analyzed. Help Analyst to outline the required insight using Mapreduce. | 6,7,8 | CO2 |
| 2 | An Electronic item manufacturing company wants to hold all the product details using hive. The data related to employees in that manufacturing company is stored in HDFS. How best we can access the details about the employees from HDFS by directly mentioning the location of the HDFS data file during the creation of employees table | 10,11 | CO3 |
| 3 | The analyst wants to know how many tweets are neutral, negative, positive based on sentiment field. Help Analyst to outline required insight using spark analytical tool | 15 | CO4 |

**Target set for course outcome attainment:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Sl.**  **No** | **C.O. No.** | **Course Outcomes** | **Target set for attainment in percentage** |
| 01 | CO1 | Describe the fundamental concepts of big data analytics | 70% |
| 02 | CO2 | Apply Map-Reduce programming on the given datasets to extract required insights | 65% |
| 03 | CO3 | Employ appropriate Hadoop Ecosystem tools such as Hive, Hbase to perform data analytics for a given problem | 65% |
| 04 | CO4 | Use Spark tool to analyze the given dataset for a given problem | 65% |



**Signature of the Course Instructor**

**This course has been duly verified Approved by the D.A.C.**

**Signature of the Chairperson D.A.C.**

**Course Completion Remarks &Self-Assessment.[*This has to be filled after the completion of the course*]**

**[Please mention about the course coverage details w.r.t. the schedule prepared and implemented. Any specific suggestions to incorporate in the course content. Any Innovative practices followed and its experience. Any specific suggestions from the students about the content, Delivery, Evaluation etc.]**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sl.**  **No.** | **Activity**  **As listed in the course Schedule** | **Scheduled Completion Date** | **Actual Completion Date** | **Remarks** |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

Any specific suggestion/Observations on content/coverage/pedagogical methods used etc.:

Course Outcome Attainment:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Sl.**  **No.** | **C.O. No.** | **Course Outcomes** | **Target set for attainment in percentage** | **Actual C.O. Attainment**  **In Percentage** | **Remarks on attainment & Measures to enhance the attainment** |
| 01 | CO1 | Describe the fundamental concepts of big data analytics | 70% | 72.62 | Target attained |
| 02 | CO2 | Apply Map-Reduce programming on the given datasets to extract required insights | 65% | 70.81 | Target attained |
| 03 | CO3 | Employ appropriate Hadoop Ecosystem tools such as Hive, Hbase to perform data analytics for a given problem | 65% | 75.30 | Target attained |
| 04 | CO4 | Use Spark tool to analyze the given dataset for a given problem | Newly added | - |  |

**Name and signature of the Course Instructor:**

**D.A.C. observation and approval:**

**BLOOM'S TAXONOMY**

Learning Outcomes Verbs at Each Bloom Taxonomy Level to be used for writing the course Outcomes.

|  |  |  |
| --- | --- | --- |
| Cognitive Level | Illustrative Verbs | Definitions |
| Knowledge | arrange, define, describe, duplicate, identify, label, list, match, memorize, name, order, outline, recognize, relate, recall, repeat, reproduce, select, state | remembering previously learned information |
| Comprehension | classify, convert, defend, discuss, distinguish, estimate, explain, express, extend, generalize, give example(s), identify, indicate, infer, locate, paraphrase, predict, recognize, rewrite, report, restate, review, select, summarize, translate | grasping the meaning of information |
| Application | apply, change, choose, compute, demonstrate, discover, dramatize, employ, illustrate, interpret, manipulate, modify, operate, practice, predict, prepare, produce, relate schedule, show, sketch, solve, use write | applying knowledge to actual situations |
| Analysis | analyze, appraise, breakdown, calculate, categorize, classify, compare, contrast, criticize, derive, diagram, differentiate, discriminate, distinguish, examine, experiment, identify, illustrate, infer, interpret, model, outline, point out, question, relate, select, separate, subdivide, test | breaking down objects or ideas into simpler parts and seeing how the parts relate and are organized |
| Synthesis | arrange, assemble, categorize, collect, combine, comply, compose, construct, create, design, develop, devise, explain, formulate, generate, plan, prepare, propose, rearrange, reconstruct, relate, reorganize, revise, rewrite, set up, summarize, synthesize, tell, write | rearranging component ideas into a new whole |
| Evaluation | appraise, argue, assess, attach, choose, compare, conclude, contrast, defend, describe, discriminate, estimate, evaluate, explain, judge, justify, interpret, relate, predict, rate, select, summarize, support, value | making judgments based on internal evidence or external criteria |