| CODE | COURSE NAME | CATEGORY | L | T | P | CREDIT |
|---|---|---|---|---|---|---|
| ITT401 | DATA ANALYTICS | PCC | 2 | 1 | 0 | 3 |

**Preamble:** This course will equip the learners with the popular technologies used in gathering, storing, manipulating, and analyzing big data. It is designed in such a way that the students will get an exposure to the analytic concepts from basic level to the advanced level.

**Prerequisites:**

- ITT201 - Data Structures
- ITT 206 - Database Management Systems
- MAT 208 - Probability, Statistics and Advanced Graph theory
- ITT 306 - Data Science

**Course Outcomes:** After completion of the course the student will be able to:

| CO No. | Course Outcome (CO) | Bloom's Category Level |
|---|---|---|
| CO 1 | Describe the introductory concepts of data analytics; integrate statistical learning into data analytic processing and tools | Level 2: Understand |
| CO 2 | Summarize the big data concepts, methods, tools and applications; explain the evolution of NoSQL with popular NoSQL products like MongoDB | Level 3: Apply |
| CO 3 | Illustrate the ideas of distributed processing with Hadoop, MapReduce paradigm and related projects namely HBase, Spark, YARN, Hive and Pig | Level 2: Understand |
| CO 4 | Experiment with R language to perform data exploration, wrangling and modelling | Level 3: Apply |
| CO 5 | Analyze how big data techniques could be used in diverse application domains of real world | Level 4: Analyze |

**Mapping of Course Outcomes with Program Outcomes**

| | PO 1 | PO 2 | PO 3 | PO 4 | PO 5 | PO 6 | PO 7 | PO 8 | PO 9 | PO 10 | PO 11 | PO 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CO 1 | 3 | 2 | 2 | 2 | - | - | - | - | - | - | - | 2 |
| CO 2 | 2 | 3 | 3 | 2 | 3 | - | - | - | - | - | - | 2 |
| CO 3 | 2 | 2 | 2 | 2 | 3 | - | - | - | - | - | - | 2 |
| CO 4 | 2 | 3 | 3 | 3 | 3 | 2 | - | - | - | 2 | - | 3 |
| CO 5 | 2 | 3 | 3 | 3 | - | 3 | 3 | - | - | 2 | - | 3 |

3/2/1: High/Medium/Low

**Assessment Pattern**

| Bloom's Category Levels | Continuous Assessment Tests | | End Semester Examination |
|---|---|---|---|
| | 1 | 2 | |
| Level 1: Remember | 10 | 10 | 20 |
| Level 2: Understand | 20 | 15 | 35 |
| Level 3: Apply | 20 | 15 | 35 |
| Level 4: Analyse | 0 | 10 | 10 |
| Level 5: Evaluate | | | |
| Level 6: Create | | | |

**Mark distribution**

| Total Marks | Continuous Internal Evaluation (CIE) | End Semester Examination (ESE) | ESE Duration |
|---|---|---|---|
| 150 | 50 | 100 | 3 hours |

**Continuous Internal Evaluation Pattern:**

| | |
|---|---|
| Attendance | : 10 marks |
| Continuous Assessment Test (2 numbers) | : 25 marks |
| Assignment/Quiz/Course project | : 15 marks |

**End Semester Examination Pattern:** There will be *two* parts; **Part A** and **Part B**. Part A contain 10 questions with 2 questions from each module, having 3 marks for each question. Students should answer *all* questions. Part B contains 2 questions from each module of which student should answer *any one*. Each question can have maximum 2 sub-divisions and carry 14 marks.

**Sample Course Level Assessment Questions**

**Course Outcome 1 (CO 1):**

1. Define data analytics.
2. Describe the different types of data analytics with examples.
3. Illustrate data analytics life cycle.
4. Explain different statistical evaluation methods or tests.

**Course Outcome 2 (CO 2):**

1. Define big data.
2. List the characteristics of big data and different technologies related to it.
3. Explain the tools NoSQL and MongoDB.
4. Explain how MongoDB can be applied to create, update, and delete documents.

**Course Outcome 3 (CO 3):**

1. Describe the HDFS framework and interface.
2. Outline the Pig and Hive architecture.
3. Illustrate the anatomy of a YARN application.
4. Compare HBase and Hive.

**Course Outcome 4 (CO 4):**

1. Explain the basic programming concepts in R.
2. Summarize how ggplot2 and dplyr are applied in visualization of R.
3. List the methods of exploratory data analysis.
4. Explore the ways of tidying data.

**Course Outcome 5 (CO 5):**

1. Discuss Recommender Systems and its types in detail with a case study of Netflix.
2. Analyze Facebook data to do a case study on citizen centric public services.
3. Analyze uplift modelling with a case study on student dropout in higher education.

**Model Question Paper**

**Course Code: ITT401**

**Course Name: Data Analytics**

**Max.Marks :100** **Duration: 3 Hrs**

**Part A**

***Answer all questions. Each question carries 3 marks (10 * 3 = 30 Marks)***

1. What is the relationship between BI and data science?
2. Differentiate between descriptive analytics and predictive analytics.
3. What are the steps involved in big data acquisition?
4. Define web data analysis.
5. Draw the architecture of Hive and explain the services provided by it.
6. How does data flow among clients that interact in HDFS?
7. What is the significance of functions gather() and spread() in tidying data? Illustrate with an example.
8. What does geom_ref_line() do? What package does it come from? Why is displaying a reference line in plots that show residuals useful and important?
9. What do you mean by hybrid filtering? What are the advantages?
10. What are the tools used in social media analytics?

## Part B

***Answer all questions. Each question carries 14 marks. (5 * 14 = 70 Marks)***

| | | | |
|---|---|---|---|
| 11 | a | With a diagram, explain the various phases of Data Analytics Lifecycle. | 10 |
| | b | What is the significance of ANOVA? | 4 |
| | | **OR** | |
| 12 | a | Describe the following resampling techniques:<br>(i)Cross-Validation (ii) Bootstrapping | 10 |
| | b | Explain any method to test the difference in sample means of two populations. | 4 |
| 13 | a | Explain the process of data pre-processing in big data acquisition. | 8 |
| | b | Write a review about moving data into and out of the database in MongoDB. | 6 |
| | | **OR** | |
| 14 | a | How is cloud computing and IoT related to big data? | 8 |
| | b | Define NoSQL. Explain Key value data stores. | 6 |
| 15 | a | Explain the role of MapReduce in Hadoop with a suitable example. | 9 |
| | b | Describe Spark with an example. | 5 |
| | | **OR** | |
| 16 | a | Explain the architecture of HDFS. Discuss on how the MapReduce framework is modified using YARN. | 7 |
| | b | Discuss on how the MapReduce framework is modified using YARN. | 7 |
| 17 | a | Define ggplot2. What are the features provided by ggplot2?What are the problems faced while using ggplot2 and how can we overcome them? | 8 |
| | b | Write the R code to import a .csv file, examine its contents and generate its descriptive statistcs | 6 |
| | | **OR** | |
| 18 | a | With examples, illustrate how these R functions help in data analysis.<br>• filter()<br>• arrange()<br>• summarize()<br>• mutate()<br>• select() | 14 |
| 19 | a | Explain the insights for using social media as a platform to improve government–citizen interaction. | 9 |
| | b | Explain different types of recommender systems. | 5 |
| | | **OR** | |
| 20 | a | Analyze uplift modelling with an appropriate example. | 7 |
| | b | Elaborate on recommender systems with Netflix application. | 7 |

## Syllabus

| |
|---|
| **Module 1: Introduction and statistics for data analytics (7 hours)** |
| Introduction and evolution of data analytics - Types of data analytics - Data analytics life cycle - Statistical methods for evaluation – Resampling |
| **Module 2: Big data, IoT, NoSQL technologies (8 hours)** |
| Introduction to big data, Related Technologies- Cloud computing, IoT, Big data generation, Big data acquisition, Big data analysis- methods and tools,Big data applications<br>Non-relational databases -MongoDB |
| **Module 3: Big data processing – Hadoop, Spark, Hive, Pig (8 hours)** |
| Hadoop, HDFS and MR, HBase, Spark, YARN, Hive, Pig |
| **Module 4:R programming for data analytics (7 hours)** |
| R programming basics for data analytics, data import and export, visualization, transformation, exploratory analysis, tidying, modelling |
| **Module 5: Popular data analytics case studies (5 hours)** |
| Recommender systems, social media analytics , churn prediction and uplift modeling with appropriate case studies |

### Text Books

1. Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data. Wiley Publishing(1st. ed.). 2015.

2. Thomas Erl, Wajid Khattak, and Paul Buhler. Big Data Fundamentals: Concepts, Drivers &Techniques. Prentice Hall Press, USA.(1st. ed.). 2016.

3. Michael Berthold and David J. Hand. Intelligent Data Analysis: An Introduction Springer-Verlag, Berlin, Heidelberg.(1st. ed.). 1999.

4. Min Chen, Shiwen Mao, Yin Zhang, and Victor C. M. Leung. Big Data: Related Technologies, Challenges and Future Prospects. Springer Publishing Company, Incorporated.2014.

5. Shashank Tiwari. Professional NoSQL.Wrox Press Ltd., GBR. 2011.

6. Kristina Chodorow and Michael Dirolf. Mongo DB: The Definitive Guide. O'Reilly Media, Inc. (1st. ed.). 2010.

7. Tom White. Hadoop: The Definitive Guide. O'Reilly Media, Inc.(4th. ed.). 2015.

8. Hadley Wickham and Garrett Grolemund. R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. O'Reilly Media, Inc.(1st. ed.). 2017.

9. Bart Baesens. Analytics in a Big Data World: The Essential Guide to Data Science and its Applications. Wiley Publishing.(1st. ed.). 2014.

**References**

1. Michael Minelli, Michele Chambers, and AmbigaDhiraj. Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses. Wiley Publishing.(Wiley CIO) (1st. ed.). 2013.

2. EelcoPlugge, Tim Hawkins, and Peter Membrey. The Definitive Guide to MongoDB: The NoSQL Database for Cloud and Desktop Computing. Apress, USA. (1st. ed.). 2010.

3. Joe Celko. Joe Celko's Complete Guide to NoSQL: What Every SQL Professional Needs to Know about Non-Relational Databases. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. (1st. ed.). 2013.

4. Benjamin Bengfort and Jenny Kim. Data Analytics with Hadoop: An Introduction for Data Scientists. O'Reilly Media, Inc. (1st. ed.). 2016.

5. Brett Lantz. Machine Learning with R. Packt Publishing. (2nd. ed.). 2015.

6. The R Manuals - https://cran.r-project.org/manuals.html

7. Carlos A. Gomez-Uribe and Neil Hunt. (2016). *The Netflix Recommender System: Algorithms, Business Value, and Innovation*. ACM Trans. Manage. Inf. Syst. 6, 4, Article 13 (January 2016), 19 pages. DOI:https://doi.org/10.1145/2843948

8. Chicago Reddick, C., Chatfield, A., &Ojo, A. (2017). *A social media text analytics framework for double-loop learning for citizen-centric public services*: A case study of a local government Facebook use. Gov. Inf. Q., 34, 110-125.

9. Diego Olaya, Jonathan Vásquez, Sebastián Maldonado, Jaime Miranda, WouterVerbeke, *Uplift Modeling for preventing student dropout in higher education*,Decision Support Systems,Volume 134, 2020,113320, ISSN 0167-9236,https://doi.org/10.1016/j.dss.2020.113320.

**Course Contents and Lecture Schedule**

| Sl. No. | Topic | No. of Lectures |
|---|---|---|
| **1** | **Introduction and statistics for data analytics** | **7 Hours** |
| 1.1 | Introduction and evolution of data analytics (Text1: 1.1, 1.1.2, 1.2) | 1 |
| 1.2 | Data Analytics Lifecycle (Text1: 2.1 -2.7) | 1 |
| 1.3 | Types of data analytics (descriptive, prescriptive, predictive, diagnostic) (Text2: 1) | 1 |
| 1.4 | Statistical Methods for Evaluation (Text1: 3.3) | 2 |

| | | |
|---|---|---|
| 1.5 | Resampling (Text3: 2.6) | 2 |
| **2** | **Big data, IoT, NoSQL technologies** | **8 Hours** |
| 2.1 | Introduction to big data-Definition, features and challenges (Text4:Ch.1) | 1 |
| 2.2 | Related Technologies-Cloud computing and IoT(Text4:Ch.2- 2.1,2.2) | 1 |
| 2.3 | Big data Generation  and Acquisition(Text4:Ch.3 – 3.1,3.2) | 1 |
| 2.4 | Big data analysis - (Text4:Ch.5 - 5.2, 5.3, 5.4) | 1 |
| 2.5 | Big data applications (Text4:Ch.6 - 6.2) | 1 |
| 2.6 | NoSQL:introduction and need for NoSQL, column oriented stores, key-value stores, document databases and graph databases (Text5:Ch.1) | 1 |
| 2.7 | MongoDB features , database, collection, documents, data types, configuration, shell,(Text6:Ch.1, 2) | 1 |
| 2.8 | Creating, updating, and deleting documents ,Querying (Text6:Ch.3,4) | 1 |
| **3** | **Big data processing – Hadoop, Spark, Hive, Pig** | **8 Hours** |
| 3.1 | What is Hadoop, brief history of Hadoop, comparison with other systems (Text7:Ch.1) | 1 |
| 3.2 | MapReduce data flow, weather dataset example (Text7:Ch.2) | 1 |
| 3.3 | Hadoop Distributed File System (HDFS) concepts, basic commands, HDFS Java interface (Text7:Ch. 3) | 1 |
| 3.4 | HBase (Text7:Ch.17) | 1 |
| 3.5 | YARN, anatomy of a YARN application, scheduling (Text7:Ch. 4) | 1 |
| 3.6 | Pig Latin language, running an example, comparison with databases (Text7:Ch. 16) | 1 |
| 3.7 | Hive data warehousing, shell, running an example, Hive architecture, comparison with databases (Text7:Ch. 17) | 1 |
| 3.8 | Spark framework, example, anatomy of a SPARK job run (Text7:Ch.19) | 1 |
| **4** | **R programming for data analytics** | **7 Hours** |
| 4.1 | R programming: basics (Text8: Ch.1) | 1 |
| 4.2 | Data visualization with ggplot2 (Text8: Ch.1) | 1 |
| 4.3 | Data transformation with dplyr (Text8: Ch.3) | 1 |
| 4.4 | Exploratory data analysis in R (Text8: Ch.5) | 1.5 |
| 4.5 | Tidy data with tidyr (Text8: Ch.9) | 1.5 |
| 4.6 | Modelling (Text8: Ch. 18) | 1 |
| **5** | **Popular data analytics case studies** | **5 Hours** |

| | | |
|---|---|---|
| 5.1 | Recommender system, types ( Text9: Ch.8) | 1 |
| 5.2 | Case study: Netflix Recommender system (Ref.7) | 1 |
| 5.3 | Social media analytics: current trends, tools (Text9: Ch.8) | 1 |
| 5.4 | Social media analytics for citizen-centric public services: a case study of a local government Facebook use (Ref.8) | 1 |
| 5.5 | Churn prediction (Text9: Ch.8)<br>Uplift modelling Case study: Uplift Modeling for preventing student dropout in higher education (Ref.9) | 1 |

| CODE | COURSE NAME | CATEGORY | L | T | P | CREDIT |
|---|---|---|---|---|---|---|
| ITL411 | DATA ANALYTICS LAB | PCC | 0 | 0 | 3 | 2 |

**Preamble:** Data analytics lab is a practical course to supplement the Data analytics theory course. The implementation of machine learning algorithms using R and experimenting with the dynamic, interactive visualization techniques using Tableau will equip the students to pursue careers in the data analytics domain. A familiarization of the popular analytic tools like Hadoop can help in academic projects or to carry out data analysis in new application areas.

**Prerequisites:**

- ITT201 - Data Structures
- ITT 206 - Database Management Systems
- MAT 208 - Probability, Statistics and Advanced Graph theory
- ITT 306 - Data Science

**Course Outcomes:** After the completion of the course the student will be able to:

| CO No. | Course Outcome (CO) | Bloom's Category Level |
|---|---|---|
| CO 1 | Solve simple problems of statistical analysis of data using Microsoft Excel | Level 3: Apply |
| CO 2 | Analyze the textual dataand time series data with the data visualization techniques in R | Level 3: Analyze |
| CO 3 | Implement the basic statistical techniques and machine learning algorithms using R | Level 3: Apply |
| CO 4 | Execute HDFS commands and apply Map Reduce technologies associated with big data analytics using HADOOP | Level 3: Apply |
| CO 5 | Analyzereal world data by applying the suitable visualization techniques in Tableau | Level 4: Analyze |

**Mapping of Course Outcomes with Program Outcomes**

| | PO 1 | PO 2 | PO 3 | PO 4 | PO 5 | PO 6 | PO 7 | PO 8 | PO 9 | PO 10 | PO 11 | PO 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **CO 1** | 3 | 3 | 3 | 3 | 3 | - | - | - | - | - | - | 2 |
| **CO 2** | 3 | 3 | 3 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | - | 2 |
| **CO 3** | 3 | 3 | 3 | 2 | 3 | - | - | - | - | - | - | 2 |
| **CO 4** | 3 | 3 | 3 | 2 | 3 | - | - | - | - | - | - | 2 |
| **CO 5** | 3 | 3 | 3 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | - | 2 |

3/2/1: High/Medium/Low