Lingnan University

Department of Computing and Decision Sciences

Course Syllabus

Course Title	:	Big Data Analytics
Course Code	:	CDS4005
Recommended Study Year	:	4
No. of Credits/Term	:	3
Mode of Tuition	:	Sectional Approach
Class Contact Hours	:	3 hours per week
Category in Major Prog.	:	Elective
Prerequisite(s)	:	CDS3004 Data Mining
Co-requisite	:	Nil
Exclusion	:	Nil
Exemption Requirement	:	Nil

Brief Course Description:

This course provides an understanding of the concept and challenge of big data. The focus is on the data analytic techniques to tack the V's (volume, velocity, variety, veracity and value) in big data and how these impacts data collection, monitoring, storage, analysis and reporting. Apache Hadoop and Spark are examples of big data management systems to manage and process large-scale data. The following topics across the big data domain will be introduced: distributed file systems; similarity search techniques; high-performance processing algorithms for data streams; big data search and query technology. Big data analytics applications in data science will be elaborated. Students will actively participate in the delivery of this course through assignments, portfolio development, and projects.

Aims:

This course aims to provide an advance technology and skills for students to demonstrate mastery of data collection, processing, analysis, retrieval, mining, visualization, and prediction. Students could synthesize methods from information retrieval, statistical data analysis, data mining, machine learning, and other big-data related fields. In alignment with best industry practices, students will be expected to work in a fast-paced, collaborative environment and to demonstrate independence and leadership. Students must be able to create and use tools to handle very large transactional, text, network, behavioral, and/or multimedia data sets.

Learning Outcomes (LOs):

On completion of this course, students will be able to:

1. Identify and use computational, mathematical, statistical, and modeling methods in big data analytics (PLO5);
2. Store and manage big data from different sources properly (PLO6);
3. Process and analyze big data (PLO6);
4. Visualize data and develop a report of the results (PLO6);
5. Evaluate and validate the results obtained in the big data analytic process (PLO7); and
6. Interpret the results and deliver the findings (PLO8).

Indicative Contents:

Big Data Characteristics and Concepts

5V's (volume, velocity, variety, veracity, and value)

Data Structure

Data Processing and "Big Data" (Extract, Transform, Load)

Data and Data Science Capability as a Strategic Asset

Business Problems and Data Science Solutions

Business Intelligence & Big Data

Big Data Adoption & Planning Considerations

From Business Problems to Data Mining Tasks

Other Analytics Techniques and Technologies

Big Data Storage and Processing

Master-slave Versus Peer-to-Peer Replication

CAP Theory (Consistency, Availability and Partition Tolerance)

ACID (Atomicity, Consistency, Isolation, Durability) Versus BASE (Basically Available, Soft State, Eventual Consistency)

Parallel and Distributed Data Processing

The Hadoop Ecosystem

Introduction to Hadoop

Hadoop components: MapReduce/Blocks/YARN (Yet Another Resource Manager)

Hadoop File System Shell

Data Snapshot in Hadoop

The Spark Ecosystem

Introduction to Spark

Resilient Distributed Dataset

Spark Operators

Processing and Analyzing Data in Spark

Data Analytics in the Context of Big Data

Supervised, Unsupervised, Self-supervised Methods

Models, Induction, and Prediction

Visualizing Segmentations

Representing and Mining Text

Text Representation (e.g., Word2vec, BERT)

Text Similarity (e.g., Word Mover Distance, Hamming Distance, Levenshtein Distance)

Text Sequence Modelling (e.g., Sequence-to-Sequence Model, ROUGE Evaluation)

Cross-modality Modelling (e.g., speech-to-text, image-to-text)

Graph Analytics

Introduction of Graph

Graph structures (diameter, connectivity, centrality)

Basic graph statistics

Visual analysis of graphs

Teaching Method:

There are different teaching and learning activities including lecture and laboratory sections. The concepts and principles of big data analytic will be discussed in the lecture sections. The practical skills and techniques for big data analytics will be taught during the laboratory sections. Students are required to perform a group project to apply the concepts and principles covered in this course to critically analyze the given problem(s) and creatively formulate the solution(s). Students implement the big data analytic solution(s) using relevant methods, and deliver a management report.

Measurement of Learning Outcomes:

	Class Attendance and Participation	Assignments	Group Project	Examination
1. Identify and use computational, mathematical, statistical, and modeling methods in big data analytics (PLO5)	X	X	X	X
2. Store and manage big data from different sources properly (PLO6)		X	X	
3. Process and analyze big data (PLO6)		X	X	X
4. Visualize data and develop a report of the results (PLO6)			X	
5. Evaluate and validate the results obtained in big data analytic process (PLO7)	X		X	X
6. Interpret the results and deliver the findings (PLO8)	X		X	X

1. There are a number of classroom activities to evaluate whether the students can recognize and use computational, mathematical, statistical, and modeling methods in big data analytics. Students are expected to evaluate and validate the results obtained in the big data analytic process; students are expected to interpret the results and deliver the findings (LO1, LO5, LO6, PLO5, PLO7, and PLO8).
2. The assignments require students to recognize and use computational, mathematical, statistical, and modeling methods in big data analytics. These evaluate whether students can store and manage big data from different sources and process and analyze big data (LO1, LO2. LO3, and PLO6).
3. The Group Project evaluates whether the students can recognize and use computational, mathematical, statistical, and modeling methods in big data analytics. It assesses whether students can store and manage big data from different sources and process and analyze big data. Students are expected to visualize data and develop a report of the results, and interpret the results and deliver the findings. Students are expected to be able to evaluate and validate the results obtained in the big data analytic process. (LO1-6, and PLO5-8).
4. The examination requires students to recognize and use computational, mathematical, statistical, and modeling methods in big data analytics. This can evaluate whether students can store and manage big data from different sources and process and analyze big data. Students are expected to interpret the results and deliver the findings. The ability of students to evaluate and validate the results obtained in big data analytic process is assessed in the examination (LO1, LO3, LO5-6 and PLO5-8).

Assessment:

Class Attendance and Participation	5%
Assignments	20%
Group Project	25%
Examination	50%
Total	100%

Required/Essential Readings:

1. Provost, F., and T. Fawcett, T. *Data Science For Business*. Sebastopol, CA: O'Reilly, 2013.
2. North, M. *Data Mining For the Masses*. United States: CreateSpace Independent Publishing Platform, 2016
3. Brath, Richard and David Jonker. *Graph Analysis and Visualization: Discovering Business Opportunity in Linked Data*, Wiley, 2015.
4. Lusher, Dean, Johan Koskinen, and Garry Robins. *Exponential Random Graph Models for Social Networks: Theory, Methods, and Applications*. Cambridge University Press, 2012.
5. deRoos, Dirk. *Hadoop For Dummies*. Wiley, 2014.

Recommended/Supplementary Readings:

1. EMC Education Services. *Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, 1st Edition*. John Wiley & Sons, 2015.
2. Bengfort, Benjamin and Jenny Kim. *Data Analytics with Hadoop*. O'Reilly, 2016.
3. Knaflic, Cole Nussbaumer. *Storytelling with Data: A Data Visualization Guide for Business Professionals*. Wiley, 2015.
4. White, Tom. *Hadoop: The Definitive Guide*. O'Reilly, 2015.
5. Marz, Nathan and James Warren. *Big Data: Principles and Best Practices of Scalable Realtime Data Systems, 1st edition*. Manning Publications, 2015.
6. Sumit Gupta, Shilpi. *Real-time Big Data Analytics*. Packt Publishing, 2016.
7. Mayer-Schönberger, Viktor and Kenneth Cukier, *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Houghton Mifflin Harcourt, 2013.
8. Hurwitz, Judith, Alan Nugent, Fern Halper, and Marcia Kaufman. *Big Data For Dummies*. For Dummies, 2013.
9. Bahga, Arshdeep and Vijay Madisetti. *Big Data Science & Analytics: A Hands-On Approach, 1st Edition*. VPT, 2016.
10. Evergreen, Stephanie. *Effective Data Visualization: The Right Chart for the Right Data, 1st Edition*. SAGE Publications, Inc., 2016.

Important Notes:

(1) Students are expected to spend a total of 9 hours (i.e. 3 hours of class contact and 6 hours of personal study) per week to achieve the course learning outcomes.

(2) Students shall be aware of the University regulations about dishonest practice in course work, tests and examinations, and the possible consequences as stipulated in the Regulations Governing University Examinations. In particular, plagiarism, being a kind of dishonest practice, is "the presentation of another person's work without proper acknowledgement of the source, including exact phrases, or summarised ideas, or even footnotes/citations, whether protected by copyright or not, as the student's own work". Students are required to strictly follow university regulations governing academic integrity and honesty.

(3) Students are required to submit writing assignment(s) using Turnitin.
(4) To enhance students' understanding of plagiarism, a mini-course "Online Tutorial on Plagiarism Awareness" is available on https://pla.ln.edu.hk/.

Rubric for Examination of CDS4005 - Big Data Analytics

Criteria	Very good	Satisfactory	Unsatisfactory
Identify and use computational, mathematical, statistical, and modeling methods in big data analytics	Demonstrates a clear understanding of computational, mathematical, statistical, and modeling methods in big data analytics. Correctly uses most methods to solve big data analytic problems. Most methods and results are correct.	Demonstrates an understanding of computational, mathematical, statistical, and modeling methods in big data analytics. Correctly uses the many methods to solve big data analytic problems, but some methods and results are incorrect.	Demonstrates a clear understanding of computational, mathematical, statistical, and modeling methods in big data analytics. Correctly uses the methods to solve most big data analytic problems. However, most methods and results are incorrect.
Process and analyze big data	Correctly processes and analyzes most data problems by using a data management system. Most results and methods are correct.	Correctly processes and analyzes some data problems by using a data management system. Some results and methods are correct.	Correctly processes and analyzes some data problems by using a data management system. However, most results and methods are incorrect.
Evaluate and validate the result obtained in big data analytic process	Identifies many methods to evaluate and validate the results obtained in a data science process.	Identifies some methods to evaluate and validate the results obtained in a data science process.	The student does not know how to evaluate and validate the results obtained in a data science process.
Interpret the results and deliver the findings	Clearly and correctly delivers most findings and communicates with different stakeholders with a variety of diverse backgrounds.	Clearly delivers many findings and communicates with different stakeholders with a variety of diverse backgrounds. However, some findings are unclear or only partially correct.	Delivers the findings and communicates with different stakeholders with a variety of diverse backgrounds. But most findings are unclear and incorrect.

Rubric for Group Project of CDS4005 - Big Data Analytics

Criteria	Very good	Satisfactory	Unsatisfactory
Identify and use computational, mathematical, statistical, and modeling methods in big data analytics	Demonstrates a clear understanding on computational, mathematical, statistical, and modeling methods in big data analytics. Correctly uses most methods to solve big data analytic problems. Most methods and results are correct.	Demonstrates an understanding on computational, mathematical, statistical, and modeling methods in big data analytics. Correctly uses the many methods to solve big data analytic problems, but some methods and results are incorrect.	Demonstrates a clear understanding on computational, mathematical, statistical, and modeling methods in big data analytics. Correctly uses the methods to solve most big data analytic problems. However, most methods and results are incorrect.
Store and manage big data from different sources	Correctly stores many formats of data to a database management system.	Correctly stores some formats of data to a database management system. However, several formats of data are applied using a wrong method.	Incorrectly stores many formats of data to a database management system. However, several formats of data are applied using a correct method.
Process and analyze big data	Correctly processes and analyzes most data problems by using a data management system. Most results and methods are correct.	Correctly processes and analyzes some data problems by using a data management system. Some results and methods are correct.	Correctly processes and analyzes some data problems by using a data management system. However, most results and methods are incorrect.
Visualize data and develop the data report of results	Demonstrates many methods to visualize data, and develops a comprehensive data report of the results. Most methods and results are correct.	Demonstrates some methods to visualize data, and develops a data report of partial results. Some methods and results are correct.	The student cannot demonstrate some methods to visualize data, and develops a data report of the results. Most methods and results are incorrect.
Evaluate and validate the result obtained in the big data analytic process	Identifies many methods to evaluate and validate the results obtained in a data science process.	Identifies some methods to evaluate and validate the results obtained in a data science process.	The student do not know how to evaluate and validate the results obtained in a data science process.
Interpret the results and deliver the findings	Clearly and correctly delivers most findings and communicates with different stakeholders with a variety of diverse backgrounds.	Clearly delivers many findings and communicates with different stakeholders with a variety of diverse backgrounds. However, some findings are unclear or only partially correct.	Delivers the findings and communicates with different stakeholders with a variety of diverse backgrounds. But most findings are unclear and incorrect.

Rubric for Assignments of CDS4005 - Big Data Analytics

Criteria	Very good	Satisfactory	Unsatisfactory
Identify and use computational, mathematical, statistical, and modeling methods in big data analytics	Demonstrates a clear understanding of computational, mathematical, statistical, and modeling methods in big data analytics. Correctly uses most methods to solve big data analytic problems. Most methods and results are correct.	Demonstrates an understanding of computational, mathematical, statistical, and modeling methods in big data analytics. Correctly uses some methods to solve big data analytic problems, but some methods and results are incorrect.	Demonstrates a clear understanding of computational, mathematical, statistical, and modeling methods in big data analytics. Correctly uses the methods to solve most big data analytic problems. However, most methods and results are incorrect.
Store and manage big data from different sources	Correctly stores many formats of data to a database management system.	Correctly stores some formats of data to a database management system. However, several formats of data are applied in a wrong method.	Incorrectly stores many formats of data to a database management system. However, several formats of data are applied in a correct method.
Process and analyze big data	Correctly processes and analyzes most data problems by using a data management system. Most results and methods are correct.	Correctly processes and analyzes some data problems by using a data management system. Some results and methods are correct.	Correctly processes and analyzes some data problems by using a data management system. However, most results and methods are incorrect.

Rubric for Class Attendance and Participation of Big Data Analytics

Criteria	Very good	Satisfactory	Unsatisfactory
Identify and use computational, mathematical, statistical, and modeling methods in big data analytics	Demonstrates a clear understanding on computational, mathematical, statistical, and modeling methods in big data analytics. Correctly uses most methods to solve big data analytic problems. Most methods and results are correct.	Demonstrates an understanding on computational, mathematical, statistical, and modeling methods in big data analytics. Correctly uses some methods to solve big data analytic problems, but some methods and results are incorrect.	Demonstrates a clear understanding on computational, mathematical, statistical, and modeling methods in big data analytics. Correctly uses the methods to solve most big data analytic problems. However, most methods and results are incorrect.
Evaluate and validate the results obtained in the big data analytic process	Identifies many methods to evaluate and validate the results obtained in a data science process.	Identifies some methods to evaluate and validate the results obtained in a data science process.	The student do not know how to evaluate and validate the results obtained in a data science process.
Interpret the results and deliver the findings	Clearly and correctly delivers most findings and communicates with different stakeholders with a variety of diverse backgrounds.	Clearly delivers many findings and communicates with different stakeholders with a variety of diverse backgrounds. However, some findings are unclear or only partially correct.	Delivers the findings and communicates with different stakeholders with a variety of diverse backgrounds. But most findings are unclear and incorrect.