Title (Units): **COMP7095 Big Data Management (3,2,1)**

Course Aims: To introduce the fundamental issues of big data management; To learn the numeracy skills and the latest techniques for big data management and processing; To conduct application case studies to show how data management techniques support large-scale data processing.

Prerequisite: Basic Knowledge in Database Concepts, Scripting and Programming.

Course Intended Learning Outcomes (CILOs):
Upon successful completion of this course, students should be able to:

No.	Course Intended Learning Outcomes (CILOs)
	Knowledge
1	Identify big data management problems and applications.
2	Describe data management framework and methods.
3	Describe applicability of data management applications.
	Professional Skill
4	Suggest appropriate solutions and tools to data management problems.
5	Analyze data management algorithms and techniques.
	Attitude
6	Work as a team to solve challenging data management problems.

Calendar Description: This course aims to introduce fundamental issues of big data management and the common data management techniques and tools including streaming and I/O efficient algorithms, distributed computing, and NoSQL storage and processing. In addition, advanced topics and an in-depth case study of Apache Spark for various applications on streaming, SQL, and graphs will also be covered. Furthermore, the potential applications of big data management to several areas such as business, finance, and so forth, are shown via examples and practices.

Teaching and Learning Activities (TLAs):

CILOs	Type of TLA
1-3	Students will learn the concepts and algorithms from lectures.
4-5	Students will learn the big data management tools and techniques via tutorials, assignments, and guided laboratory.
4-6	Students will work on a group project to gain hands-on experience of big data management.

Assessment:

No.	Assessment Methods	Weighting	CILOs to be addressed	Description of Assessment Tasks
1	Continuous Assessment	40%	4-6	Assignments and Labs will be used to consolidate their knowledge and develop their skills in data management. Group project will further strengthen their understanding and problem-solving skills.
2	Examination	60%	1-5	Final Examination questions are designed to evaluate how far students have achieved their intended learning outcomes. Analytics based questions will be used to assess the understanding of big data management and processing problems. Problem solving questions will be used to assess the students' ability in tackling applications in big data management.

Assessment Rubrics:

	Excellent (A)	Good (B)	Satisfactory (C)	Fail (F)
Identify and distinguish the problems of big data management and analytics	Thorough identification of almost all problems	Identification of a large number of problems	Identification of a moderate number of problems	Identification of very small number of problems
Describe big data management algorithms	Thorough description of almost all big data management algorithms	Description of most of the algorithms	Description of some of the algorithms	Description of only a few number of algorithms
Describe applicability of big data management	Thorough description of almost all usage of big data management	Description of most of the usage	Description of some of the usage	Description of very small number of usage
Suggest appropriate solutions to data management problems	Suggestion of almost all correct solutions	Suggestion of most of the solutions	Suggestion of some of the solutions	Suggestion of very small number of solutions
Analyze big data management methodologies and techniques	Thorough analysis of almost all data management methodologies and techniques	Analysis of most of the methodologies and techniques	Analysis of some methodologies algorithms and techniques	Analysis of very small number of methodologies and techniques

Course Content and CILOs Mapping:

Content		CILO No.
I	Introduction to Big Data	1
II	Big Data Platforms	2,3,5
III	NoSQL Storage and Processing	2,3,5
IV	Big Data Summarization and Visualization	3,4,5,6
V	Advanced Topics	3,4,5,6
VI	An in-depth case study of Apache Spark for various applications on streaming, SQL, machine learning, and graphs	3,4,5,6

References:

- Ghavami, Peter. "Big Data Management: Data Governance Principles for Big Data Analytics". Walter de Gruyter GmbH & Co KG, 2020.Thomas Erl,Wajid Khattak, and Paul Buhler, "Big Data Fundamentals: Concepts, Drivers & Techniques" (1st edition), Pearson, ISBN-13: 978-0134291079, 2016
- Li, Kuan-Ching, Hai Jiang, and Albert Y. Zomaya. "Big data management and processing". CRC Press, 2017.
- Miller, James D. "Big data visualization". Packt Publishing Ltd, 2017.
- Karau, Holden, and Rachel Warren. "High performance Spark: best practices for scaling and optimizing Apache Spark". O'Reilly Media, Inc., 2017.
- Sandy Ryza, Uri Laserson, Sean Owen, Josh Wills, "Advanced Analytics with Spark" (2nd edition), ISBN: 9781491972908, 2017.

Course Content:

Topic

I. Introduction to Big Data

II. Big Data Platforms
   A. Hadoop
   - Hadoop Distributed File System
   - MapReduce
   B. Spark (RDD)

III. NoSQL Storage and Processing
   A. NoSQL Data Management: Matrix Data, Graph Data, Text Data, and Image/Video Data Management
   B. Case studies: MongoDB

IV. Big Data Summarization and Visualization

V. Advanced Topics
   A. Streaming Algorithms and Applications
   B. Distributed Algorithms and Platforms

VI. An in-depth case study of Apache Spark for various applications on streaming, SQL, machine learning, and graphs
   A. Spark Dataframe and Spark SQL
   B. Spark MLlib
   C. Spark Streaming
   D. Spark GraphX