Subject Description Form

Subject Code	DSAI4205
Subject Title	Big Data Analytics
Credit Value	3
Level	4
Pre-requisite / Co-requisite / Exclusion	**Pre-requisites**: AMA1104/COMP1004, and COMP1011/COMP1012/EIE1003/ENG2002, and COMP2011/COMP2013/DSAI2201/EIE3320, and COMP2411/EIE3112, or equivalent introductory database subject
Objectives	The objectives of this subject are to: 1. introduce students the concept and challenge of big data (3 V's: volume, velocity, and variety); and 2. teach students in applying skills and tools to manage and analyse the big data.
Intended Learning Outcomes	Upon completion of the subject, students will be able to: (a) understand the concept and challenge of big data and why existing technology is inadequate to analyse the big data; (b) understand how to collect, manage, store, and query various form of big data; (c) understand how to analyse big data using various quantitative methods; (d) gain hands-on experience on large-scale analytics tools to solve some open big data problems; and (e) be able to conduct thorough analysis on the impact of big data for business decisions and strategy in real-world applications.
Subject Synopsis/ Indicative Syllabus	**Topic** **1. Introduction to Big Data** Different V's, their challenges and application domains. **2. Collection of Big Data** Eventual Consistency and NoSQL systems (MongoDB, BigTable, etc.) **3. Large-Scale Data Analytics Systems** Hadoop, MapReduce, Hive, etc. **4. Basic Statistical Analysis** **5. Machine Learning Systems for Big Data** **6. Graph Analytics** Graph structures, PageRank, Centrality, etc.

	7. Data Analysis Application: Recommender System **8. Data Visualisation**
Teaching/ Learning Methodology	A mix of lectures and lab sessions is used to deliver the various topics in this subject. Lectures are conducted to initiate students with the concepts and techniques of big data. Students are given the opportunity to gain hands-on experience on both open- source and commercial big data analytics software during the laboratory sessions.

Assessment Methods in Alignment with Intended Learning Outcomes

Specific assessment methods/tasks	% weighting	Intended subject learning outcomes to be assessed				
		a	b	c	d	e
Continuous Assessment						
1. Lab Exercises / Assignments	**60%**	✓	✓	✓	✓	✓
2. Project		✓	✓	✓	✓	✓
3. Quiz		✓	✓	✓		
Examination	**40%**	✓	✓	✓		✓
Total	100 %					

Explanation of the appropriateness of the assessment methods in assessing the intended learning outcomes:

Continuous assessments consist of a project, assignments, lab exercises, and quizzes, which are designed to facilitate students to achieve intended learning outcomes. Lab exercise is designed to encourage students to acquire deep understanding of the relevant knowledge, practice in order to enrich their hands-on experience with various software tools. The project is designed to enhance students' ability to acquire the understanding and using different knowledge, principles, techniques, tools to solve a real problem through team. Quizzes are to ensure the students understand the concepts.

Examination will evaluate student's understanding and usage of big data technologies.

Student Study Effort Expected		
	Class contact:	
	▪ Lecture/Tutorial/Laboratory	39 Hrs.
	Other student study effort:	
	▪ Self-study and Revisions	62 Hrs.
	▪ Project	15 Hrs.
	Total student study effort	116 Hrs.

Reading List and References	**Reference Books:** 1. Segaran, Toby, and Jeff Hammerbacher, *Beautiful data: the stories behind elegant data solution*, O'Reilly Media, Inc., 2009 2. Dean, Jeffrey and Ghemawat, Sanjay, "MapReduce: simplified data processing on large clusters", *Communications of the ACM*, January 2008. 3. Stonebraker, M., Abadi, D., DeWitt, David J., Madden, S., Paulson, E., Pavlo, A. and Rasin, A., "MapReduce and Parallel DBMS's: Friends or Foes?", *Communications of the ACM*, January 2010. 4. Dean, Jeffrey and Ghemawat, Sanjay, "MapReduce: A Flexible Data Processing Tool", *Communications of the ACM*, January 2010. 5. K. Shvachko, H. Kuang, S. Radia and R. Chansler, "The Hadoop Distributed File System", *IEEE Symposium on Mass Storage Systems and Technologies*, 2010 6. White, Tom, *Hadoop: The definitive guide*, O'Reilly Media, Inc., 2012. 7. Cattell, Rick, "Scalable SQL and NoSQL Data Stores", *ACM SIGMOD Record*, Volume 39, Issue 4, December 2010. 8. Page, Lawrence and Brin, Sergey and Motwani, Rajeev and Winograd, Terry, "The PageRank Citation Ranking: Bringing Order to the Web", Technical Report, Stanford InfoLab, 1999. 9. Toby Segaran, *Programming Collective Intelligence*, O'Reilly Media, Inc., 2007 10. Han, Jiawei, Micheline Kamber, and Jian Pei, *Data Mining: Concepts and Techniques*, 3rd Edition, Morgan Kauffman, 2011. 11. Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar, *Introduction to data mining*, Pearson Education India, 2016. 12. Murphy, Kevin P., *Machine Learning: A Probabilistic Perspective*, MIT press, 2012. 13. Theobald, Oliver, *Machine Learning for Absolute Beginners: A Plain English Introduction*, 2nd Edition, Scatterplot press, 2017. 14. The NumPy community, NumPy: The Absolute Basics for Beginners. 15. The Pandas community, 10 minutes to Pandas 16. Géron, A., Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems, O'Reilly Media, 2019 17. Turnbull, J. The Docker Book: Containerization is the new virtualization, James Turnbull, 2014.