Subject Description Form

Subject Code	COMP4434
Subject Title	Big Data Analytics
Credit Value	3
Level	4
Pre-requisite / Co-requisite / Exclusion	**Pre-requisites**: AMA1104, COMP1011, COMP2011, COMP2411
Objectives	The objectives of this subject are to: • introduce students the concept and challenge of big data (3 V’s: volume, velocity, and variety); and • teach students in applying skills and tools to manage and analyze 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 analyze the big data; (b) collect, manage, store, query, and analyze various form of big data; (c) gain hands-on experience on large-scale analytics tools to solve some open big data problems; and (d) understand the impact of big data for business decisions and strategy.
Subject Synopsis/ Indicative Syllabus	**Topic** **1. Introduction to Big Data** The 3 V’s, their challenges and application domains. **2. Collection of Big Data** Eventual Consistency and NoSQL systems MongoDB, Google BigTable. **3. Large-Scale Data Analytics Systems** Auto-Parallel Data Programming; MapReduce, Hive, and Parallel Databases **4. Basic Statistical Analysis** Fruad and Benfords Law, Bayesian Introduction, Heteroskedasticity **5. Machine Learning Systems for Big Data** **6. Graph Analytics** Graph structures (diameter, connectivity, centrality), PageRank, Triangle counting

7. Sentiment Analysis

8. Data Visualization

Data types and dimensions; Visual encoding and perception

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 (Please tick as appropriate)			
		a	b	c	d
Continuous Assessment	**60%**				
1. Lab Exercises / Assignments		✓	✓	✓	✓
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:	
▪ Lectures	26 Hrs.
▪ Tutorials/Laboratory	13 Hrs.
Other student study effort:	
▪ Review the lecture	28 Hrs.
▪ Review the lab	14 Hrs.

	▪ Work on the project	15 Hrs.
	▪ Prepare the quizzes	9 Hrs.
	▪ Prepare the examination	11 Hrs.
	Total student study effort	116 Hrs.
Reading List and References	**Reference Books:** 1. Dolan, J.C.B., Dunlap, M., Hellerstein, J.M. and Welton, C., *MAD Skills: New Analysis Practices for Big Data*, 2009. 2. Rajaraman, Anand and Ullman, Jeffery David, *Mining of Massive Datasets*, Chapters 1-2, 2011. 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. Lin, Jimmy and Dyer, Chris, *Data-Intensive Text Processing with MapReduce*, Morgan and Claypool, 2010. 6. Cattell, Rick, "Scalable SQL and NoSQL Data Stores", *ACM SIGMOD Record*, Volume 39, Issue 4, December 2010. 7. Elmagarmid, Ahmed K., Ipeirotis, Panagiotis G. and Verykios, Vassilios S., "Duplicate Record Detection: A Survey", *IEEE Transactions on Knowledge and Data Engineering*, Volume 19, Issue 1, January 2007. 8. Koudas, N., Sarawagi, S. and Srivastava, D., "Record Linkage: Similarity Measures and Algorithms", *Proceedings of the ACM SIGMOD International Conference on Management of Data*, June 2006. 9. Hothorn, Torsten and Everitt, Brian S., *A Handbook of Statistical Analyses Using R*, 3rd Edition, Chapter 3, CRC Press, 2014. 10. Gregory Park on overfitting to the leaderboard in a Kaggle Competition. 11. Wu, X.D., Kumar, V., Quinlan, J. Ross, Ghosh, J., Yang, Q. and et al., "Top 10 Algorithms in Data Mining, Knowledge and Information Systems", *Journal of knowledge and Information Systems*, Volume 14, Issue 1, page 1-37, 2007. (Read C4.5) 12. Domingos, Pedro, "A Few Useful Things to Know about Machine Learning", *Communications of the ACM*, Volume 55, Issue 10, 2012. 13. Alpaydin, Ethem, *Introduction to Machine Learning*, 3rd Edition, MIT Press, 2015. 14. Haykin, Simon, *Neural Networks and Learning Machines*, 3rd Edition, Pearson, 2016.	

	15. Hanaran, Pat, Tools for Data Enthusiasts. 16. Heer, J., Bostock, M. and Ogievetsky, V., “A Tour through the Visualization Zoo”, *Communications of the ACM*, Volume 53, Issue 6, June 2010.