Subject Description Form

Subject Code	COMP4434
Subject Title	Big Data Analytics
Credit Value	3
Level	4
Pre-requisites	AMA1104 Introductory Probability COMP1011 Programming Fundamentals COMP2011 Data Structures COMP2411 Database Systems
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). 2. 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; and (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	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

<table>
<tr><td></td><td>8. Data Visualization: Data types and dimensions; Visual encoding and perception</td></tr>
<tr><td>Teaching/Learning Methodology</td><td>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.</td></tr>
<tr><td>Assessment Methods in Alignment with Intended Learning Outcomes</td><td>
<table>
<tr><td rowspan="2">Specific assessment methods/tasks</td><td rowspan="2">% weighting</td><td colspan="6">Intended subject learning outcomes to be assessed (Please tick as appropriate)</td></tr>
<tr><td>a</td><td>b</td><td>c</td><td>d</td><td></td><td></td></tr>
<tr><td>1. Lab exercises/Assignments</td><td rowspan="3">60%</td><td>x</td><td>x</td><td>x</td><td>x</td><td></td><td></td></tr>
<tr><td>2. Project</td><td>x</td><td>x</td><td>x</td><td>x</td><td></td><td></td></tr>
<tr><td>3. Quiz</td><td>x</td><td>x</td><td></td><td></td><td></td><td></td></tr>
<tr><td>4. Examination</td><td>40%</td><td>x</td><td>x</td><td></td><td>x</td><td></td><td></td></tr>
<tr><td>Total</td><td>100 %</td><td colspan="6"></td></tr>
</table>

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.</td></tr>
<tr><td>Student Study Effort Expected</td><td>Class contact:</td><td></td></tr>
<tr><td></td><td>▪ Lecture</td><td>26 Hrs.</td></tr>
<tr><td></td><td>▪ Tutorial/Laboratory</td><td>13 Hrs.</td></tr>
<tr><td></td><td>Other student study effort:</td><td></td></tr>
</table>

	▪ 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	1. How Vertica Was the Star of the Obama Campaign, and Other Revelations 2. Cohen et al.“MAD Skills: New Analysis Practices for Big Data”, 2009 3. Ullman, Rajaraman, Mining of Massive Datasets, Chapter 2 4. Stonebraker et al., “MapReduce and Parallel DBMS’s: Friends or Foes?”, Communications of the ACM, January 2010. 5. Dean and Ghemawat, “MapReduce: A Flexible Data Processing Tool”, Communications of the ACM, January 2010. 6. Rick Cattell, “Scalable SQL and NoSQL Data Stores”, SIGMOD Record, December 2010 (39:4) 7. Elmagarmid, et. al. "Duplicate Record Detection: A Survey" 8. Koudas, et. al. "Record Linkage: Similarity Measures and Algorithms" 9. Chapter 3 of A Handbook of Statistical Analyses Using R 10. Gregory Park on overfitting to the leaderboard in a Kaggle Competition 11. Xindong Wu et al., Top 10 Algorithms in Data Mining, Knowledge and Information Systems, 14(2008), 1: 1-37. (read C4.5) 12. Ullman, Rajaraman, Mining of Massive Datasets , Chapter 1 13. Pedro Domingos, A Few Useful Things to Know about Machine Learning, CACM 55(10), 2012 14. Pat Hanaran, Tools for Data Enthusiasts 15. Jeffrey Heer, Michael Bostock, Vadim Ogievetsky, A Tour through the Visualization Zoo, Communications of the ACM, Volume 53 Issue 6, June 2010 16. Howard Wen, "Big Ethics for Big Data", O'Reilly Media	