Subject Description Form

Subject Code	COMP5434
Subject Title	Big Data Computing
Credit Value	3
Level	5
Pre-requisites	Knowledge in database systems, machine learning and data analytics is preferred.
Objectives	The objectives of this subject are to: 1. introduce students the concept and challenge of big data; 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 traditional technology is inadequate to analyze the big data; b. understand how to collect, manage, store, and query various form of big data; c. familiar with the classical data analysis and machine learning algorithms; d. familiar with large-scale analytics tools to solve some open big data problems; and e. analyze the impact of big data for real-world business decisions and strategy.
Subject Synopsis/ Indicative Syllabus	1. Introduction to Big Data: Different V's, their challenges and application domains. 2. Cloud Computing Basics: Software as a service (SaaS), Platform as a Service (PaaS), Infrastructure as a Service (IaaS), Desktop as a Service (DaaS), Public, Private and Enterprise Cloud. 3. Big Data Computing: Concepts, Platform, Service, and Tools 4. Large-Scale Programming Abstraction: MapReduce and its open source implementation of Hadoop 5. Large-Scale Data Processing Framework: Apache Spark and its Built-in Modules 6. Large-Scale Database Management: NoSQL and other tools, e.g. MongoDB, Google BigTable, etc. 7. Machine Learning Systems for Big Data: Methods and Tools 8. Big Data Visualization: Data types and dimensions; Visual encoding and perception 9. Big Data Case Studies

<table>
<tr><td>Teaching/Learning Methodology</td><td>A mix of lectures, discussions and case studies.

Class activities include lectures, tutorials, laboratory works and seminars.</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="5">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>e</td></tr>
<tr><td>1. Assignments or lab works</td><td rowspan="3">55</td><td>✓</td><td>✓</td><td>✓</td><td>✓</td><td>✓</td></tr>
<tr><td>2. Project</td><td>✓</td><td>✓</td><td>✓</td><td>✓</td><td>✓</td></tr>
<tr><td>3. Quiz</td><td>✓</td><td>✓</td><td>✓</td><td>✓</td><td></td></tr>
<tr><td>4. Examination</td><td>45</td><td>✓</td><td>✓</td><td>✓</td><td></td><td>✓</td></tr>
<tr><td>Total</td><td>100</td><td colspan="5"></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 good 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>
<table>
<tr><td>Class contact:</td><td></td></tr>
<tr><td>Class activities (lecture, tutorial, lab, etc.)</td><td>39 Hrs.</td></tr>
<tr><td>Other student study effort:</td><td></td></tr>
<tr><td>Assignments, Quizzes, Projects, Examination</td><td>65 Hrs.</td></tr>
<tr><td>Total student study effort</td><td>104 Hrs.</td></tr>
</table>
</td></tr>
<tr><td>Reading List and References</td><td>

Jared Dean, Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners. Wiley, 2014.
Steele, Julie, and Noah Iliinsky, Beautiful visualization: looking at data through the eyes of experts, O'Reilly Media, Inc., 2010.
Dean, Jeffrey and Ghemawat, Sanjay, "MapReduce: simplified data processing on large clusters", Communications of the ACM, January 2008.
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.

</td></tr>
</table>

5. Dean, Jeffrey and Ghemawat, Sanjay, “MapReduce: A Flexible Data Processing Tool”, Communications of the ACM, January 2010.
6. Lin, Jimmy and Dyer, Chris, Data-Intensive Text Processing with MapReduce, Morgan and Claypool, 2010.
7. K. Shvachko, H. Kuang, S. Radia and R. Chansler, “The Hadoop Distributed File System”, IEEE Symposium on Mass Storage Systems and Technologies, 2010.
8. White, Tom, Hadoop: The definitive guide, O'Reilly Media, Inc., 2012.
9. Cattell, Rick, “Scalable SQL and NoSQL Data Stores”, ACM SIGMOD Record, Volume 39, Issue 4, December 2010.
10. Chodorow, Kristina. MongoDB: the definitive guide: powerful and scalable data storage, O'Reilly Media, Inc., 2013.
11. Silberschatz, Abraham, Henry F. Korth, and Shashank Sudarshan, Database System Concepts, 7th Edition, 2019.
12. 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.
13. Wu, X.D., Kumar, V., Quinlan, J. Ross, Ghosh, J., Yang, Q. 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.
14. Leskovec, Rajaraman, Ullman, Mining of Massive Datasets, 2nd Edition, Cambridge University Press, 2014.
15. Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar, Introduction to data mining, Pearson Education India, 2016.
16. Hastie, Trevor, Robert Tibshirani, and Jerome Friedman, The Elements of Statistical Learning: Data mining, Inference, and Prediction, Springer Science & Business Media, 2009.
17. Bishop, Christopher M., Pattern Recognition and Machine Learning, Springer, 2006.
18. Goodfellow, Ian, et al., Deep Learning: Adaptive Computation and Machine Learning series, MIT press, 2016.
19. McKinney, W., Python for data analysis: Data wrangling with Pandas, NumPy, and IPython, O'Reilly Media, Inc., 2012.
20. Hothorn, Torsten and Everitt, Brian S., A Handbook of Statistical Analyses Using R, CRC Press, 2014.
21. Géron, A., Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems, O'Reilly Media, 2019.
22. Nickoloff, J., Docker in action, Manning Publications Co., 2016.