H7SDA: Scalable Data Analytics

Module Code:	H7SDA
Long Title	Scalable Data Analytics APPROVED
Title	Scalable Data Analytics
Module Level:	LEVEL 7
EQF Level:	6
EHEA Level:	First Cycle

Credits:	5

Module Coordinator:	Horacio Gonzalez-Velez

Module Author:	Horacio Gonzalez-Velez

Departments:	School of Computing

Specifications of the qualifications and experience required of staff	MSc and/or PhD degree in computer science or cognate discipline. May have industry experience also.

Learning Outcomes

On successful completion of this module the learner will be able to:

#	Learning Outcome Description
LO1	Describe and apply MapReduce and extensions for creating parallel applications on large amounts of data
LO2	Describe and summarise search techniques including similarity search and search engine technologies.
LO3	Distinguish between data-stream processing and specialised algorithms
LO4	Develop analytical and ethical skills to employ mining and clustering algorithms on large multi-dimensional datasets

Dependencies

Module Recommendations

No recommendations listed

Co-requisite Modules

No Co-requisite modules listed

Entry requirements	Learners should have attained the knowledge, skills and competence gained from stage 2 of the BSc (Hons) in Data Science

H7SDA: Scalable Data Analytics

Module Content & Assessment

Indicative Content
MapReduce I Definition of the MapReduce paradigm
MapReduce II Algorithms using MapReduce
MapReduce Extensions Recursive and workflow systems for MapReduce. Resilient data sets.
MapReduce Cost Models Complexity and cost models for MapReduce with emphasis on communication costs and task networks
Near Neighbour search and Shingling Collaborative filtering and similarity sets. Document shingling and sub-strings.
Hashing Locality-sensitive hashing and distance measures. Additional methods for higher degrees of similarity.
Stream Data Model Stream sources, stream queries, and processing. Sampling data
Streams Operations I Filtering and counting.
Streams Operations II Combining and estimating
Page Rank PageRank algorithm in its application to search engines. Efficient computation of PageRank.
Link Analysis Link Spam. Hubs and authorities.
Clustering Techniques Points, spaces and distances. Dimensionality.

Assessment Breakdown	%
Coursework	50.00%
End of Module Assessment	50.00%

Assessments

Full Time

Coursework

Assessment Type:	Continuous Assessment	**% of total:**	Non-Marked
Assessment Date:	n/a	**Outcome addressed:**	1,2,3,4
Non-Marked:	Yes		

Assessment Description:
Ongoing feedback on ongoing tutorial activities. Feedback on regular reflection.

Assessment Type:	Continuous Assessment	**% of total:**	50
Assessment Date:	n/a	**Outcome addressed:**	1,4
Non-Marked:	No		

Assessment Description:
This practical assessment will evaluate the learners’ knowledge and understanding of scalable data analytics, possibly in the context of MapReduce, mining and/or clustering algorithms. A marking scheme is provided in Appendices.

Assessment Type:	Easter Examination	**% of total:**	50
Assessment Date:	n/a	**Outcome addressed:**	2,3
Non-Marked:	No		

Assessment Description:
The test will assess learners’ knowledge and understanding of search and stream processing techniques. A sample question, marking scheme, and solution, is provided in Appendices.

No End of Module Assessment

No Workplace Assessment

Reassessment Requirement

Repeat examination
Reassessment of this module will consist of a repeat examination. It is possible that there will also be a requirement to be reassessed in a coursework element.

Reassessment Description
The repeat strategy for this module is a terminal assessment. Students will be afforded an opportunity to repeat the assessment at specified times throughout the year and all learning outcomes will be assessed in the repeat assessment.

H7SDA: Scalable Data Analytics

Module Workload

Module Target Workload Hours 0 Hours

Workload: Full Time

Workload Type	*Workload Description*	*Hours*	*Frequency*	*Average Weekly Learner Workload*
Lecture	Classroom & Demonstrations (hours)	24	Per Semester	2.00
Tutorial	Other hours (Practical/Tutorial)	12	Per Semester	1.00
Independent Learning	Independent learning (hours)	89	Per Semester	7.42
			Total Weekly Contact Hours	3.00

Module Resources

Recommended Book Resources

Leskovec, J., Rajaraman, A. & Ullman, J.D.. (2014), Mining of Massive Datasets (2nd ed), Cambridge University Press.

Kleppmann, M.. (2017), Designing Data-Intensive Applications: The Big Ideas behind Reliable, Scalable, and Maintainable Systems, O'Reilly Media.

Kolodziej, J. & González-Vélez, H.. (2019), High-Performance Modelling and Simulation for Big Data Applications, Springer International Publishing.

Supplementary Book Resources

Marz, N. & Warren, J.. (2015), Big Data: Principles and best practices of scalable real-time data systems, Manning Publications.

White, T.. (2015), Hadoop: The Definitive Guide (4th ed), O'Reilly Media.

McCool, M., Reinders, J. & Robison, A.D.. (2012), Structured Parallel Programming: Patterns for Efficient Computation, Morgan Kaufmann.

Holmes, A.. (2014), Hadoop in Practice (2nd ed), Manning Publications.

Lublinsky, B Smith, K. T. & Yakubovich, A.. (2013), Professional Hadoop Solutions, Wrox.

Ojeda, T., Murphy, S.P. & Bengfort, B.. (2014), Practical Data Science Cookbook, Packt Publishing.

This module does not have any article/paper resources

Other Resources

Dean, J. & Ghemawat, S. (2010). MapReduce: a flexible data processing tool. Commun. ACM 53(1): 72-77..

Kolodziej, J., González-Vélez, H. & Karatza, H.D. (2017). High-performance modelling and simulation for big data applications. Simulation Modelling Practice and Theory 76: 1-2 (2017)..

Ubarhande, V., Popescu , A-M., & González-Vélez, H. (2015). Novel Data-Distribution Technique for Hadoop in Heterogeneous Cloud Environments. CISIS 2015: 217-224..

Petcu, D. et al. (2014). Next Generation HPC Clouds: A View for Large-Scale Scientific and Data-Intensive Applications. Euro-Par Workshops (2): 26-37..

González-Vélez, H., & Kontagora, M. (2011). Performance evaluation of MapReduce using full virtualisation on a departmental cloud. Applied Mathematics and Computer Science 21(2): 275-284..

Discussion Note:	