

H9SSP: Scalable Systems Programming

Module Code:	H9SSP
Long Title	Scalable Systems Programming APPROVED
Title	Scalable Systems Programming
Module Level:	LEVEL 9
EQF Level:	7
EHEA Level:	Second Cycle
Credits:	5
Module Coordinator:	Horacio Gonzalez-Velez
Module Author:	Margarete Silva
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	
<i>On successful completion of this module the learner will be able to:</i>	
#	Learning Outcome Description
LO1	Demonstrate in-depth knowledge of parallel algorithms on large amounts of data
LO2	Identify and categorise search techniques including similarity search and search engine technologies.
LO3	Critically compare and contrast different data-stream processing and specialised algorithms.
LO4	Critically analyse mining and clustering algorithms on large multi-dimensional datasets.
LO5	Develop and implement efficient programming solutions for problems relating to processing data at scale.
Dependencies	
Module Recommendations	
No recommendations listed	
Co-requisite Modules	
No Co-requisite modules listed	
Entry requirements	A level 8 degree or its equivalent in any discipline

H9SSP: Scalable Systems Programming

Module Content & Assessment			
Indicative Content			
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 Filtering, counting, combining and estimating.			
Stream Processing Building complex pipelines and models			
Link Analysis PageRank algorithm in its application to search engines. Efficient computation of PageRank. Link Spam. Hubs and authorities.			
Frequent itemsets Market-Basket Model, many-to-many relationships. Association rules.			
A-Priori / Limited Pass Algorithms Determine stages, sets and items under different monotonicity conditions.			
Clusters for Streams and Parallelism Bucket initialisation and merging. Parallel clustering.			
Using Scalable Services Deploying concurrent stream processing and batch processing pipelines			
Assessment Breakdown			%
Coursework			100.00%
Assessments			
Full Time			
Coursework			
Assessment Type:	Formative Assessment	% of total:	Non-Marked
Assessment Date:	n/a	Outcome addressed:	1,2,3,4,5
Non-Marked:	Yes		
Assessment Description: Formative assessment will be provided on the in-class individual or group activities. Feedback will be provided in written or oral format, or on-line through Moodle. In addition, in class discussions will be undertaken as part of the practical approach to learning.			
Assessment Type:	Continuous Assessment	% of total:	100
Assessment Date:	n/a	Outcome addressed:	1,2,3,4,5
Non-Marked:	No		
Assessment Description: This practical assessment will evaluate the learners' knowledge and understanding of Scalable Systems Programming, possibly in the context of mining and/or clustering algorithms.			
No End of Module Assessment			
No Workplace Assessment			
Reassessment Requirement			
Coursework Only <i>This module is reassessed solely on the basis of re-submitted coursework. There is no repeat written examination.</i>			
Reassessment Description Reassessment of this module will be via project			

H9SSP: Scalable Systems Programming

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	Every Week	24.00
Tutorial	Other hours (Practical/Tutorial)	24	Every Week	24.00
Independent Learning	Independent learning (hours)	77	Every Week	77.00
Total Weekly Contact Hours				48.00

Module Resources	
<i>Recommended Book Resources</i>	
<p>Jure Leskovec, Anand Rajaraman, Jeffrey David Ullman. (2014), Mining of Massive Datasets, Cambridge University Press, p.476, [ISBN: 1107077230].</p> <p>Martin Kleppmann. (2017), Designing Data-intensive Applications, O'Reilly & Associates Incorporated, p.590, [ISBN: 1449373321].</p>	
<i>Supplementary Book Resources</i>	
<p>Andrew Kelleher, Adam Kelleher. (2018), Machine Learning in Production, [ISBN: 9780134116556].</p> <p>K. Hwang. (2017), Cloud and Cognitive Computing: A Machine Learning Approach, MIT Press, [ISBN: 10026203641X].</p> <p>B. Chambers, M. Zaharia. (2018), Spark - The Definitive Guide, 1st. O'Reilly Media, [ISBN: 101491912219].</p> <p>Tom White. Hadoop: the Definitive Guide ; Storage and Analysis at Internet Scale, [ISBN: 1491901632].</p>	
<i>Recommended Article/Paper Resources</i>	
<p>B. Veloso, F. Leal, H. González-Vélez, B. Malheiro, J-C. Burguillo. (2018), Scalable data analytics using crowdsourced repositories and streams, Journal of Parallel and Distributed Computing, 122, p.1-10.</p> <p>J. Eckroth. (2018), A course on big data analytics, Journal of Parallel and Distributed Computing, 118, p.166.</p> <p>J. Kołodziej, H. González-Vélez, H.D. Karatza. (2017), High-performance modelling and simulation for big data applications, Simulation Modelling Practice and Theory, 76, p.1-2.</p> <p>J. Dean, S. Ghemawat. (2010), MapReduce: a flexible data processing tool., Commun. ACM, 53(1), p.72-77.</p>	
<i>This module does not have any other resources</i>	
Discussion Note:	