6.6 Big Data Storage and Processing

6.6.1 Headline information about the module

Module title	Big Data Storage and Processing
Module NFQ level (only if an NFQ level can be demonstrated)	9
Module number/reference	M6
Parent programme(s) the plural arises if there are embedded programmes to be validated.	MSc in Data Analytics
Stage of parent programme	AWARD
Semester (semester1/semester2 if applicable)	2
Module credit units (FET/HET/ECTS)	ECTS
Module credit number of units	10
List the teaching and learning modes	Full time & Part time blended learning
Entry requirements (statement of knowledge, skill and competence)	See Section 4
Pre-requisite module titles	none
Co-requisite module titles	none
Is this a capstone module? (Yes or No)	No
Specification of the qualifications (academic, pedagogical and professional/occupational) and experience required of staff (staff includes workplace personnel who are responsible for learners such as apprentices, trainees and learners in clinical placements)	Academic and Professional: PhD desirable and a minimum of an MSc is required. However, in exceptional cases, NFQ Level 8 in Data Analytics, Computer Science, Software Development, Software Engineering or equivalent may be acceptable when combined with significant industrial experience. Pedagogical: Teaching experience is desired. Completion of postgraduate CPD/Certificate in Teaching and Learning or similar preferred. Experience in blended learning delivery required. In absence of experience, training will be mandatory and will be provided.
Maximum number of learners per centre (or instance of the module)	120
Duration of the module	1 semester
Average (over the duration of the module) of the contact hours per week (see * below)	5 Hours
Module-specific physical resources and support required per centre (or instance of the module)	Physical resource requirements are 1 laptop or PC/workstation per student. On campus and online resources as per programme specification.
Analysis of required learning effort	
*Effort while in contact with staff	

On campus Lecture / Classroom demonstrations		Mentoring and small-group tutoring		Other (Reflective development, directed reading/group work)		Online classes & Directed e-learning (hours)	Independent learning (hours)	Other hours (specify)	Work-based learning hours of learning effort	Total effort (hours)
Hours	Minimum ratio teacher/learner	Hours	Minimum ratio teacher/learner	Hours	Minimum ratio teacher/learner					
10	1:60	10	1:30	10	1:60	30	190	N/A	N/A	250
Allocation of marks (within the module)										
				Continuous assessment		Supervised project	Proctored practical examination	Proctored written examination	Total	
Percentage contribution				100%					100%	

6.6.2 Module aims and objectives

The aims of this module:

1. Fundamentals of big data storage and data management paradigms
2. Underlying principles of parallel and distributed computing
3. Current solutions for retrieving, integrating and processing Big Data
4. Big Data programming models and their efficient usage at scales
5. Big Data Streams and their processing techniques

6.6.3 Minimum intended module learning outcomes

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

1. Critically assess the data storage and management requirements of a given data project from a modern perspective and evaluate limitations of legacy approaches to Big Data. (Linked to PLO 3)

2. Assess the design concepts and architectural patterns of distributed Big Data systems and analyse the components that form their technology stack. (Linked to PLO 1, PLO 2)

3. Critically evaluate and select a Big Data environment suitable for retrieving and processing a given Big Data set, perform data management operations and select appropriate analytic algorithms for the required scale and speed. (Linked to PLO 2, PLO 3)

4. Assess the functional differences between common big data environments and the particularities

of big graphs and appraise Graph Big Data management and processing stacks. (Linked to PLO 4)

5. Implement the tools and technologies that facilitate the processing of Streaming Big Data to perform real-time analytics. (Linked to PLO 2)

6.6.4 Rationale for inclusion of the module in the programme and its contribution to the overall MIPLOs

This module deals with the core enabler for Data Analytics, i.e., data. Companies generate large amounts of data that need to be gathered and stored for eventual analysis to turn them into value. This module will equip students with the analytical and technical skills to manage large and diverse amounts of data to allow their analysis at the right scale and within the right time frame.

Knowledge gained of statistical concepts, implementation and visualisation from earlier modules provides learners with the insight necessary to interpret results from large scale data analysis.

6.6.5 Information provided to learners about the module

A copy of the Module Descriptor will be provided to learners at the start of the module via the College LMS (Moodle)

6.6.6 Module content, organisation and structure

The following indicative syllabus contains a learner reflective component as outlined in the teaching and learning strategy for this programme.

Content
Legacy Approaches • Traditional Computing Architecture & Data Storage • Relational DBMS(SQL) & Data Silos • Old Data (SQL) vs. Big Data
Distributed Systems and Data Management • Architectures • Methodologies • Scaling
Big Data Storage • Physical Storage • Data Processing ETL/ELT • Data Tiering • File Formats, Compression and Security • Disaster Recovery
No-SQL • Key–value (e.g., Couchbase, Redis) • Document (e.g., MongoDB, CouchDB) • Columnar (e.g., Big Table, Cassandra) • Graph (e.g., Neo4j) • Spatial (e.g., OGC-compliant)

Big Data Platforms • Apache Hadoop and HDFS • MapReduce • YARN (resource management) • Apache Spark
Big Data Programming • Apache Hive (SQL-like queries) • Apache Pig (high-level scripts that run on Apache Hadoop) • Apache Mahout (machine learning algorithms on Apache Hadoop) • Spark MLlib (scalable and easy machine learning library on Apache Spark)
Streaming Big Data • Spark Streaming • Kafka
Graph Big Data • Apache Giraph (Graph processing on Graph Big Data)

The above topics (where applicable) will be performed within a testing environment (programming I.D.E / spreadsheet to allow practical integration of theoretical knowledge)

Programming skills are continuously developed throughout this module through practical implementation of module content. Conceptual / theoretical topics will also be practically demonstrated through the use of appropriate tool sets, to emphasise the synergy between theory and programmatic demonstration.

6.6.7 Module teaching and learning (including formative assessment) strategy

To provide the learner with a strong foundation in the core topics covered during the lectures, practical sessions will reinforce lecture content and provide supervised time to complete some assessment tasks. Sessions will be interactive, with instructor-led example exercises highlighting important topics discussed in lectures.

To provide formative assessment for this module the learner will:

- Be provided an opportunity at the beginning of each week to engage in group discussion on the material covered the previous week, thereby allowing reflection and ensuring their competency.
- Complete student-suggested tasks in a peer learning environment to encourage collaboration and allow learners to self-evaluate their current knowledge while gaining new knowledge and insights (this strategy links directly to PLO 7, PLO 8)
- Join additional discussions covering any lab-based exercises which have been provided to the learner.

Online and on campus learning activities according to learning type.

Learning Type	Online activities	On campus activities
Knowledge Acquisition	• Pre-recorded presentations / demonstrations • Multi-media text-based materials • Videos • Guest speakers • Open ed resources	• Face to face lecturers • Practical Demonstrations
Collaboration	• Group projects • Discussion forums • Virtual classroom peer learning • Team virtual lab activities • Group presentations	• Group projects • Team based lab activities / practical workshops • Group presentations

	• Mentoring	
Discussion	• Discussion forums (synchronous and asynchronous) • Zoom breakout room discussions • Online tutorials • Project supervision • Webinars (industry experts) • Reflective activities	• Class discussion • Tutorials • Project supervision • Face to face lab / practical activities
Investigation	• Open ed resources • Lab observations • Project research • Information and data sourcing, analysis and evaluation • Flipped Learning	• data sourcing and analysis
Practice	• Virtual lab • Simulations • Case studies • Analysis of data sets • Presentations • Online quizzes / MCQs	• Labs • Practical workshops • Group work
Production	• E-portfolio • Reflective journal • Assessment outputs • Quiz / MCQs • Case studies • GitHub records	• Exam • Case studies • Student demonstration

6.6.8 Work-based learning and practice-placement

Not Applicable

6.6.9 E-learning

Collaborative blended learning strategies will be utilised for this module to ensure peer learning can be experienced not only through face-to-face traditional learning but also through online approaches. The online element will be achieved through a variety of interactive methods, including discussion forums, collaborative blogs and wikis, virtual labs and classrooms, group online supervision, interactive demonstrations and e-portfolios. This integrated learning approach ensures learning can be both reflective and collaborative while developing the efficacy of the individual student as the module progresses.

6.6.10 Module physical resource requirements

Physical resource requirements are 1 laptop or PC/workstation per student. On campus and online resources as per programme specification.

6.6.11 Reading lists and other information resources

Recommended:

Bill Chambers, 2018, 1st Edition, Spark: The Definitive Guide: Big Data Processing Made Simple, O'Reilly Media, ISBN: 978-1491912218

Tom White, 2015, 4th Edition, Hadoop: The Definitive Guide: Storage and Analysis at Internet Scale, O'Reilly Media, ISBN: 978-1491901632
Tom Davenport, 2014, Big Data at Work: Dispelling the Myths, Uncovering the Opportunities, Harvard Business Review Press, ISBN: 978-1422168165

6.6.12 Module summative assessment strategy

Online and on campus assessment activities according to assessment type.

Assessment	Online Assessment	On Campus Assessment
Knowledge Acquisition	• Pre-recorded presentations / demonstrations	• Live presentations / demonstrations
Collaboration	• Group projects • Team virtual lab activities • Group pre-recorded presentations / demonstrations	• Team lab activities • Group presentations / demonstrations
Formative	• Project (individual & group) supervision • Reflective activities	• Project (individual & group) supervision
Investigation	• Theoretical and practical project research • Problem Based Learning	• Problem Based Learning
Practical	• Virtual lab • Analysis of data sets • Online quizzes / MCQs • Technical tasks (individual and group)	• Lab • Analysis of data sets • Technical tasks (individual and group)
Production	• Artefacts • Student code repository records	• Artefacts • Student code repository records

For indicative assessment schedule see appendix 8

Mapping of summative and formative assessment: MIMLO to MIPLO

A ✓ indicates that the PLO has been formatively assessed as per the teaching and learning strategy for the programme	Module Learning Outcomes					Programme Learning Outcomes							
	1	2	3	4	5	1	2	3	4	5	6	7	8
Summative Assessment												✓	✓

6.6.13 Sample assessment materials

See appendix 9