Ulster University

COM 745
Big data & infrastructure

Module Introduction

Dr. Joe Rafferty

ulster.ac.uk

1

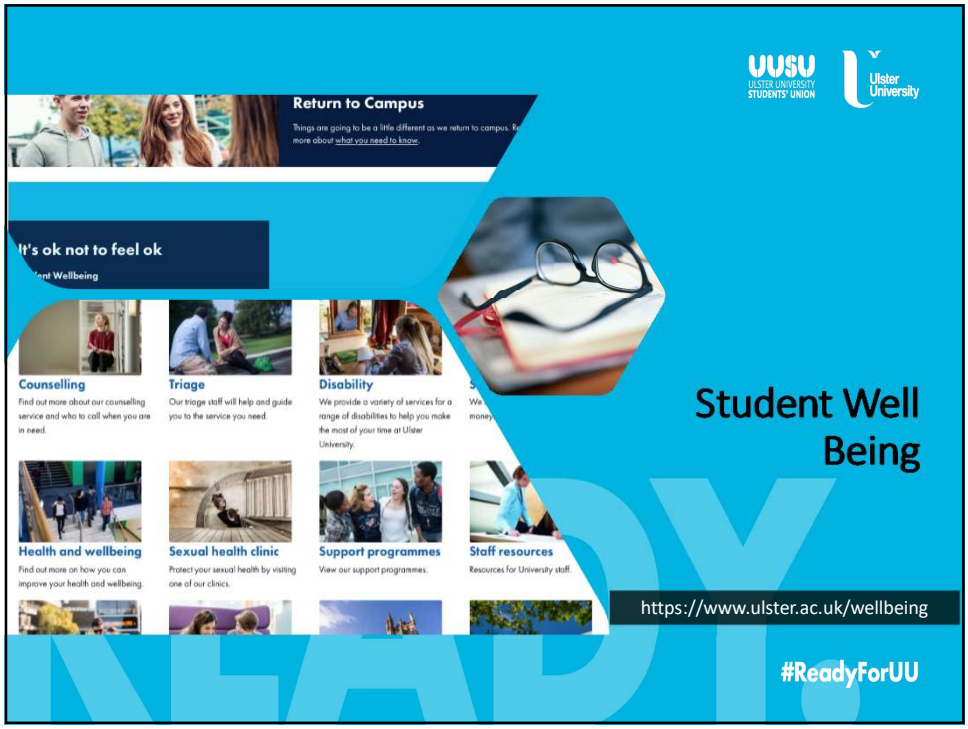
Module Overview
Module Metadata

- Module Coordinator:
  Dr Joe Rafferty
  j.rafferty@ulster.ac.uk

- Module Level – 7 (associated with PG)*
- Credit points – 20; total student effort **200** hours

Ulster University

2



3

Module Overview
Introduction

- Why this module?

4

Module Overview
Introduction

- Why does this matter for IoT?

Ulster University

5

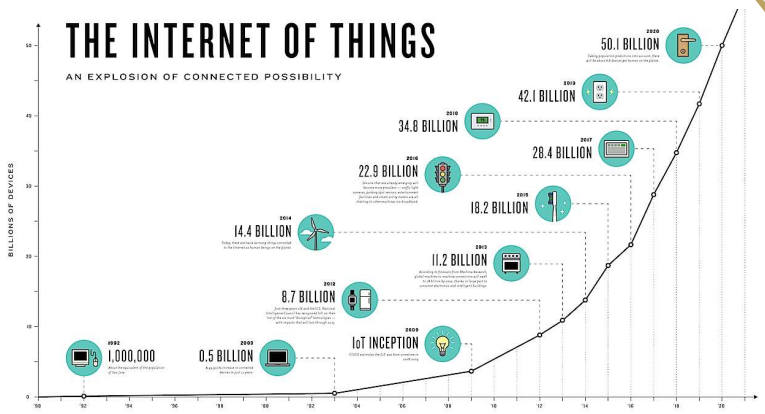
Module Overview
Introduction

- IoT solutions generate a data – which probably needs to be stored

Ulster University

6

Module Overview
Introduction



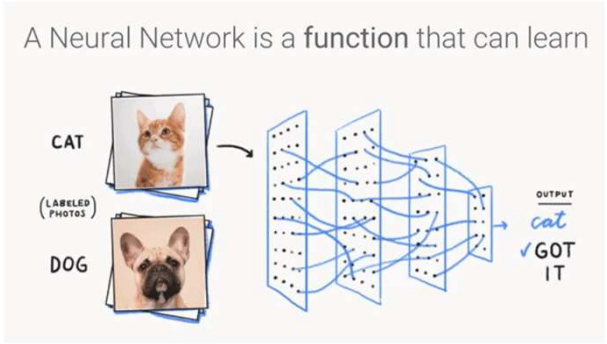
- There is an explosion in the number of IoT devices generating this data – their numbers will only grow

Ulster University

7

Module Overview
Introduction

- Why does this matter for AI?



Ulster University

8

Module Overview
Introduction

- Managing the **data generated by IoT Solutions** is a “big data” problem.
- Managing the **data required to train AI solutions** is a “big data” problem.
- Big data is defined by the three V’s;
  - high Volume of data
  - highly Variable data
  - high Velocity data

Ulster University

9

Module Overview
Introduction

- Big data employs a variety of database/data storage systems
- These database/data storage systems vary in purpose and optimisations
- Understanding of the use and purpose of these database/data storage systems will allow design and implementation of big data platforms

Ulster University

10

Module Overview
Rationale

The aims of this module are:

- Provide an understanding of the variety of database/data storage classes that may be required to realise a big data platform.
- To gain knowledge of the purpose and theory behind each class of system.
- To understand the selection criteria and use cases related to each class of system
- To provide students with the practical experience of working with a variety of database systems through graphical tools and programmatically

Ulster University

11

Module Overview
Module Content

The 3Vs problem
Introduction of the problem area, the 3 Vs

General database concepts
Function of databases
Classes of databases to be covered, including example scenarios

Ulster University

12

Module Overview
Module Content

Introduction to time series databases
Underlying concepts
Use cases
Querying, creating and manipulating data through visual tools and programming

Introduction to graph databases*
Underlying concepts
Querying, creating and manipulating data through visual tools and programming

Ulster University

13

Module Overview
Module Content

Linux command line basics
Connecting to a remote Linux server though SSH
File management
Issuing commands

Hadoop architecture and Administration
The architecture of a typical Hadoop installation, with a discussion on a range of components

Ulster University

14

Module Overview
Module Content

Introduction to functional programming concepts
Concepts behind functional programming

Introduction to map reduce
Introducing querying Hadoop data through Map reduce
Practical experience of querying data through Map reduce

Ulster University

15

Module Overview
Module Content

Introduction to Spark
Introducing querying Hadoop data through Spark
Practical experience of querying data through Spark

Introduction to Pig
Introducing querying Hadoop data through Pig
Practical experience of querying data through Pig

Ulster University

16

Module Overview
Module Content

Introduction to Hive
Understanding data warehousing using Hive
Practical management and querying of data within Hive

Introduction to Sqoop
Introduction on importing/exporting data to/from traditional SQL databases

Ulster University

17

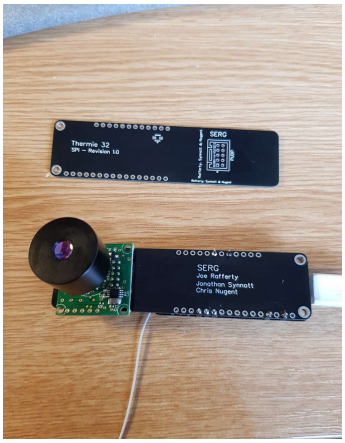
Module Overview
Module Objectives (informal)

1. To give you the background knowledge needed to select appropriate databases for your application/solution needs.
2. To provide the comprehension of underlying concepts of each type of databases.
3. Provide practical experience of using each type of database

Ulster University

18

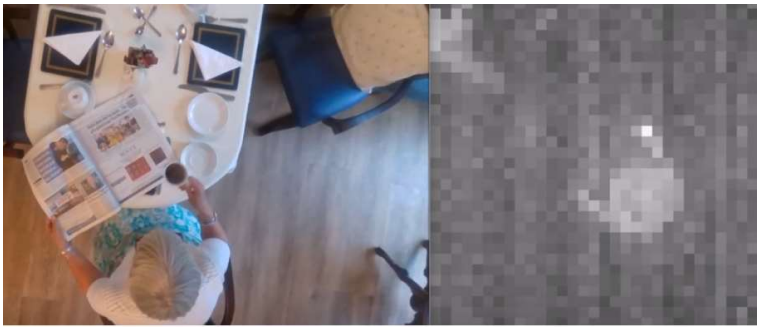
Module Overview
Case study



Ulster University

19

Module Overview
Case study

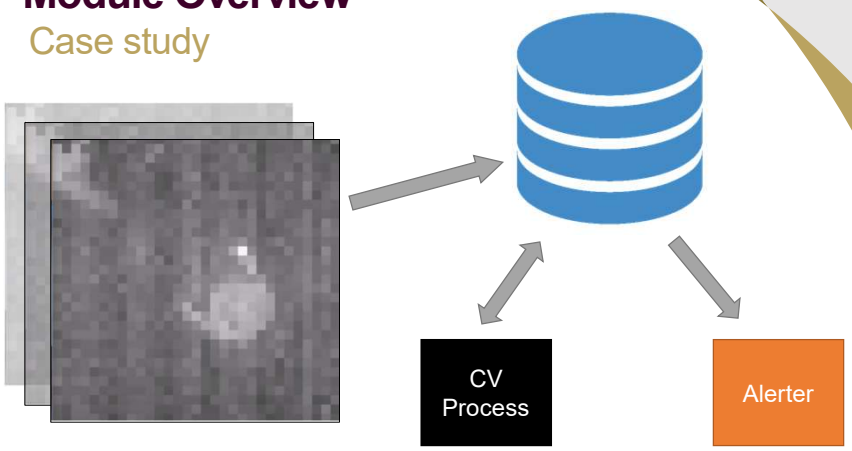


Ulster University

20

Module Overview

Case study



Ulster University

21

Module Overview

Teaching plan Week 1-5 (subject to change)

Week: Date	Topic(s)	Assessment Event
1: 25/09	Topic Background An Overview of DB technologies Relational Databases	
2: 02/10	Introduction to NoSQL Databases Document Databases	
3: 09/10	Document Databases Time Series Databases	
4: 16/10	Time Series Databases Graph Databases	CW 1 Released
5: 23/11	Graph Databases Hadoop Concepts	

Ulster University

22

Module Overview

Teaching plan Week 5-*10* (subject to change)

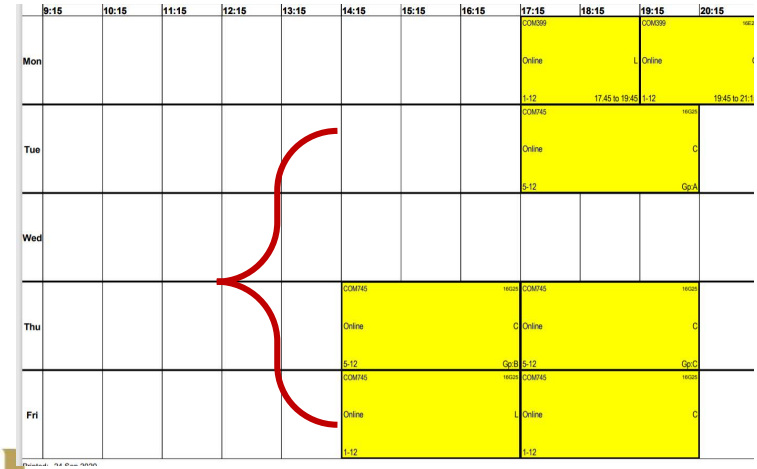
6: 30/11	Hadoop Concepts Functional Programming/MapReduce	**CW1 Deadline (17:00)**
7: 06/11	Hadoop Spark Hadoop Pig	
8: 13/11	Hadoop Hive/Sqoop*	**CW2 Released**
9: 20/11	GDPR and IoT Recap	
10: 27/11	***CW2 Support*** ***CW2 Deadline Submission @ 17:00***	

Ulster University

23

Module Overview

Weekly Timetable



Ulster University

24

Module Overview
Logistics

Week 1-5:
Lecture/Practical for **6** hours on Friday*

Week 6-10:
- Lecture for **3** hours on Friday.
- Potentially split into groups as per timetable for practical's.
- Potentially all together on Friday depending on logistics.

https://forms.office.com/Pages/ResponsePage.aspx?id=h5QLb6hPqEKutL8uLCLU6HcuWIsG-odBhmNZajXhHI5UMEEwMlpTNVMzVlQyRldETFI3REU1NUk3Qy4u

Ulster University

25

Module Overview
Assessment Plan

Assessment will be 100% in coursework based:

- Coursework 1 consists of a set exercise [50%]
- Coursework 2 is a practical skills assessment [50%]

Ulster University

26

Module Overview

Coursework 1 – A set exercise [50%]

Released: 16th October 2020
Submission Deadline: 17:00 - 30th October 2020
Feedback Date: 13th November 2020

Students are expected to design a database platform which **can store and mange sensor data for a smart city on a big data scale**.

This database platform should incorporate a variety of databases with different optimisations in order to most efficiently enable storage of the target data **at scale**

Specifically, they would like this platform to store information regarding API-keys, sensors and users.

Storage of user data will encompass:

- A unique user identifier
- an email address (which acts as a username)
- a password hash
- forename
- surname
- user right levels (1-7)
- associated organisation

API-keys will require storage of:

- the key (a string)
- issuance data
- the identifiers of the associated user

Sensor data will require storage of :

- the sensor ID
- sensor label
- sensor class
- granular location (GPS)
- coarse location (String)
- raw sensor data to be stored (however you deem appropriate)

Ulster University

27

Module Overview

It is expected that students will produce a small report covering the design of their solution. This report will be in 3 page IEEE format [2] and will be submitted through TurnitIn via Blackboard.

This report should include the following sections:

i. Abstract
A small 250-word (max) description of the problem, data to be stored, designed solution and conclusion.

ii. Problem Definition
A description of the overall problem justifying why a custom solution needs to be developed. This should critically appraise use of a single common database for such storage.

iii. Data storage analysis
This will provide an analysis of the data types to be stored and provide an assessment of a range of suitable databases. Care should be given to justify the suitability of each database through the lens of performance.

iv. Solution design
This will provide a data schema and architectural description which incorporates a graphical depiction of the overall architecture.

v. Conclusion
This will conclude describing the strengths of the design and a projection of future scalability. It is expected that any assertions regarding future scalability will be referenced.

Any assertions regarding scalability and data storage suitability must be supported by appropriate academic references. All information used in the final must be appropriately cited and referenced in IEEE style [3]. A citation manager with a word processor plugin, such as Mendeley [4] , may be used to reduce the overhead of this process.

The assessment criteria for coursework 1 is presented as an appendix to this document.

Ulster University

28

Module Overview

Appendix I – assessment criteria coursework I

	Needs improvement	Satisfactory	Excellent	Overall credit allocation
Abstract	[0%-1%] Inadequate abstract providing a rudimentary overview of the document.	[2%] Acceptable abstract, providing a moderate summary of the content and presents a synopsis of some sections.	[3%] Well-articulated abstract incorporating broad overview and a succinct synopsis of all the sections.	3%
Problem Definition	[0%-2%] Little description of the overall problem, poorly justifying why a custom solution needs to be developed. Limited critical appraisal of the use of a single common database for such storage.	[3%-5%] Moderate description of the overall problem, adequate justification of why a custom solution needs to be developed. Adequate critical appraisal of the use of a single common database for such storage.	[6%-8%] Good description of the overall problem with a good justification of why a custom solution needs to be developed. Strong critical appraisal of the use of a single common database for such storage.	8%
Data storage analysis	[0%-3%] Data schema were scrutinised with little or insufficient insight. Databases were selected with little justification.	[4%-7%] Data schema were scrutinised with adequate insight. Databases were selected with moderate justification..	[8%-10%] Data schema were scrutinised with keen insight. Databases were selected with good justification.	10%
Solution design	[0%-5%] Justification for the choice of technology applied to the problem was minimal. The solution was documented inadequately. Scalability was not justified or examined, and alternative technologies were not explored.	[6%-12%] The technology used to produce the solution was appropriate given the requirements of the workload. The solution was documented adequately incorporating of control flow diagrams. The utility of the solution was examined and justified.	[13%-20%] The technology used to produce the solution was carefully examined and logically chosen – given the requirements of the workload. Alternative technologies were examined and excluded accordingly. The solution was documented well incorporating control flows and incorporation of software architecture diagrams. Scalability of the solution was examined and justified.	20%
Conclusion	[0%-1%] Limited reflection was applied to the solution and future scalability.	[2%-3%] Meaningful reflection was applied to the solution and future scalability.	[4%-5%] Insightful reflection was applied to the solution and future scalability. Weaknesses were identified and improvements were suggested.	5%
Referencing	[0%] No referencing.	[1%] Inadequate or incorrect referencing.	[2%] Correct and appropriate referencing.	2%

Ulster University

29

Module Overview

Coursework 2 – a set exercise [50%]

Released: 13th November 2020
Submission Deadline: 17:00 – 27th November 2020
Feedback Date: 11th December 2020

The exercise will assess understanding of further concepts and demonstrate practical skills related to the **Hadoop** environment.

Students will be set an exercise where they will be expected to:

i. Identify and evaluate a number of publicly available datasets. These may be from sources such as kaggle.com or data.gov.uk
ii. Select appropriate datasets, as informed by their interests
iii. Integrate and import these datasets into a suitable data storage and processing system, providing rationale for their choice
iv. Perform meaningful analysis of the data to derive some simple useful information, as can be obtained by the dataset selected.
v. Provide visualisation of the analysis through any Hadoop-related technologies which the students deem suitable

N.B. The students are required to implement this solution using Hadoop and its related technologies.

Once the solution is produced, students are required to produce presentation which incorporates a 5-minute video capture demonstrating the solution.

It is recommended to have 10 ***content*** slides which follow the below outline:

Slide 0. Title Slide. (0%)
Slides 1 - 2. Discussion of the problem and justification of the dataset (8%).
Slides 3 - 6. Overview of the technical solution developed (12%).
Slides 7 - 8. The analysis performed, and insight obtained (10%).
Slide 9. Functionality of the recorded demonstration (15%) [5-minute video].
Slide 10. Concluding comments (3%).
Slide 11. References (2%).

Ulster University

The assessment criteria for coursework 2 is presented as an appendix to this document.

N.B. Students should be aware of the plagiarism policy of the University and submit their coursework in accordance to this.

30

Module Overview

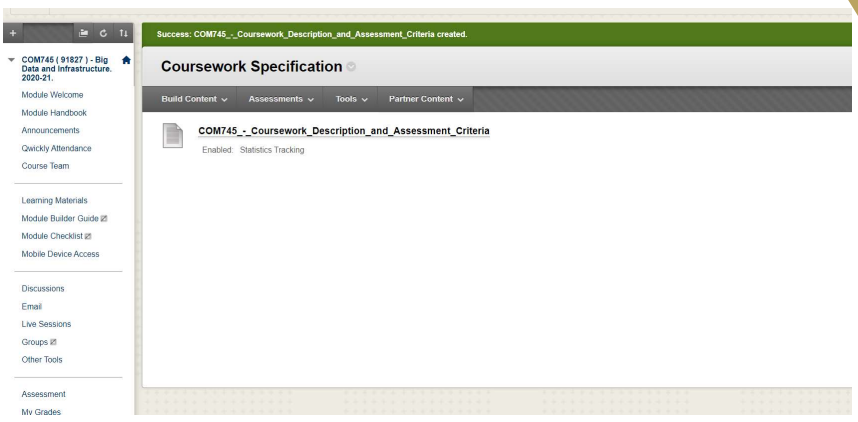
Appendix II – assessment criteria coursework 2

	Needs improvement	Satisfactory	Excellent	Overall credit allocation
Problem analysis/Selection of Dataset	[0%-2%] Datasets were selected with little or insufficient justification or reasoning. The insight that analysis of this data could provide was not elaborated upon.	[3%-5%] A problem area was faintly identified. Datasets were selected, justification may have been slight but reasonable. The avenue to insight that the datasets would provide was inferred or inadequately examined.	[6%-8%] A problem area was correctly identified, discussed and evidenced. Datasets were appropriately selected and examined. The insight that could be extracted from these datasets as reasoned upon and justified.	8%
Solution produced	[0%-4%] Justification for the choice of technology applied to the problem was minimal. The solution was documented inadequately. Scalability was not justified or examined, and alternative technologies were not explored.	[5%-8%] The technology used to produce the solution was appropriate given the requirements of the workload. The solution was documented adequately incorporating of control flow diagrams. The utility of the solution was examined and justified.	[9%-12%] The technology used to produce the solution was carefully examined and logically chosen – given the requirements of the workload. Alternative technologies were examined and excluded accordingly. The solution was documented well incorporating control flows and incorporation of software architecture diagrams. Scalability of the solution was examined and justified.	12%
Analysis/Insight from data	[0%-3%] Some analysis was attempted. Outputs may have been weak or inadequate. There may have been little or no justification of the analysis applied.	[3%-6%] A logical approach was taken to analysis. Analysis was performed across a large dataset. The analysis was adequate and some insight was provided. The analysis produced tabular output or basic graphs.	[7%-10%] A logical and informed approach to analysis was taken. Analysis was performed across a number of large datasets or a single appropriately massive dataset. The analysis was complex in nature and provided nuanced insight. The analysis produced tabular output and advanced graphical Visualisation, such as maps.	10%
Concluding comments	[0%] Limited reflection was applied to the solution and analysis.	[1%] Meaningful reflection was applied to the solution and analysis.	[2%-3%] Insightful reflection was applied to the solution and analysis. Weaknesses were identified and improvements were suggested.	3%
Referencing	[0%] No referencing.	[1%] Inadequate or incorrect referencing.	[2%] Correct and appropriate referencing.	2%
Video Demonstration: Insight offered into data	[0% - 1%] Limited metrics are produced – providing little or no insight.	[2% - 3%] Good metrics are produced - providing good insight.	[4% - 5%] Good metrics are produced in addition to incorporation of visualisation, providing advanced insight.	5%
Video Demonstration: Functionality	[0%-3%] The solution poorly or didn't operate at all.	[4%-7%] The solution functioned moderately well. Performance issues may have been present but were deemed acceptable.	[8%-10%] The solution performed well and had minimal performance issues. Advanced techniques or functionality, such as interactive visualisation, has been incorporated.	10%

31

Module Overview

Assessment Plan



32

Module Overview
Plagiarism

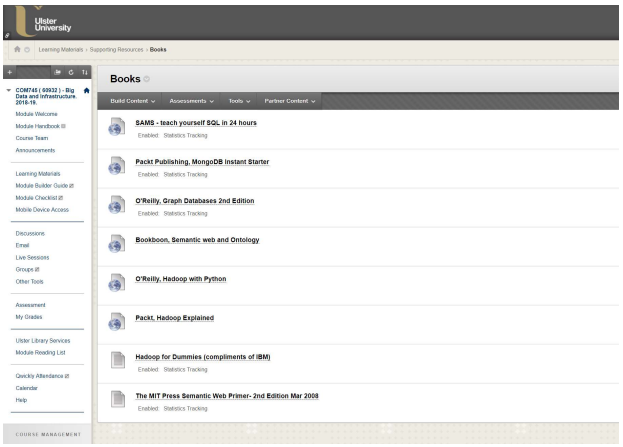
- Plagiarism is not tolerated
- It is **very** easy to detect
- Raw module stats can indicate plagiarism in isolation more so when other modules are mapped
- Machine Learning based scanning routines, such as those in Turnitin, can easily identify plagiarised work

Don't do it . Don't enable it – all parties are penalised

Ulster University

33

Module Overview
Recommended Reading



Ulster University

34

Module Overview
Required Reading

Hadoop for Dummies Special edition,
Robert D. Schneider. Wiley.
https://insidebigdata.com/white-paper/hadoop-for-dummies/ (compliments of IBM)

Ulster University

35

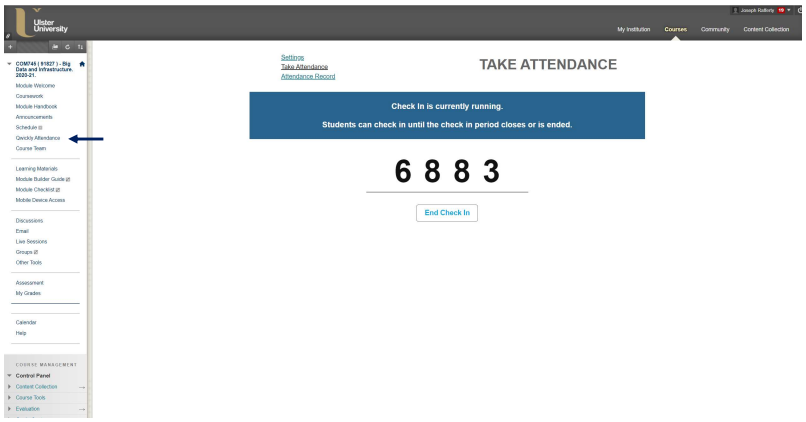
Module Overview
Recommended Reading

Hadoop with Python,
Z. Radtka, D. Miner. O' Reilly.
http://www.oreilly.com/programming/free/files/hadoop-with-python.pdf (compliments of O'Reilly)

Ulster University

36

Attendance



37

Module Overview

Any Questions?

38

39