

**6CS030 Big Data**

**2018/9**

**Portfolio**

Learning Outcomes for the Module:

LO1 – Apply appropriate theory, tools and techniques to problems associated with big data.

LO2 – Synthesise solutions to problems from the big data domain.

LO3 – Analyse and evaluate solutions to big data problems.

LO4 – Present results of solutions using appropriate methods.

The portfolio is made up of the following parts:

|  |  |  |  |
| --- | --- | --- | --- |
| **Part** | **Type** | **Learning Outcome** | **Portfolio Percentage** |
| 1 | A set of worksheet exercises based on the Workbooks and are assessed during the Workshops | LO1, LO3 | 30% |
| 2 | Coursework. Hand-in date: **12th April 2019** (Teaching Week 12) | LO2, LO3, LO4 | 40% |
| 3 | Time Constrained Assessment (TCA). To be held during Revision Week: **30th April 2019** (12 noon – to be confirmed) | LO1, LO2 | 30% |

**Passing the Module and Resits**

To pass the module you must achieve 40% overall in the portfolio. Should you not achieve a pass mark, you will be set a new piece of coursework in the resit period, to be submitted during the University’s resit week (Wednesday 10th July 2019). Note, the overall mark will be capped to 40% if you are resitting the module.

## Worksheets Resit

If you achieved less than 40% in any of the Worksheets (Part 1), or were not able to submit them there will be a final hand-in date for these on the 7th May.

Do note unless you have mitigation the mark for each worksheet submitted on the 7th May will be capped at 40%.

**Teamwork**

You can work in groups of 2 for the Coursework (Part 2). Everything else is individual work.

# REQUIREMENTS:

## Part 1 – 30 Marks

During Weeks 3, 7 and 10 you will be presented with a Worksheet exercise that should be completed during the Workshop session (hand-in by the following week’s workshop). These will be based on the previous week’s workbooks.

These are individual exercises.

## Part 2 – 40 Marks

The University currently uses relational databases for their business systems, such as e:Vision and HR. The Vice-Chancellor has heard about the Big Data Module and wants to find out more about how the techniques might benefit the University. The VC wants to commission you to write a report for the University’s IT Department that evaluates how and where Big Data Tools should be used and what for.

For example, to use externally found data to:

* Help predict future trends in student numbers
* Find out what is being said about the University

You are presented with several datasets surrounding the University and related data that could influence future student numbers:

|  |  |
| --- | --- |
| 1 | *SFR 56/2017: Key Stage 5 Destination Measures 2015/16 (revised)*  *SFR56\_2017\_KS5\_Institution\_Tables\_1516\_Revised.xlsx*  Obtained from the Department of Education. At the time of writing the latest available data was for 2015/6. Further information about the datasets can be found here:  [*https://www.gov.uk/government/statistics/destinations-of-ks4-and-ks5-pupils-2016*](https://www.gov.uk/government/statistics/destinations-of-ks4-and-ks5-pupils-2016)  *Note, the original data was in Open Document Format (ODS) and converted to Excel*  Generated on the 11th February 2019 |
| 2 | *Office for National Statistics Nomis Population Projections for years 2020-2022*  *nomis\_Population\_Projections\_2019\_02\_12.xlsx*  Downloaded from here: <https://www.nomisweb.co.uk/query/select/getdatasetbytheme.asp?opt=3&theme=&subgrp>=  Which includes data for Local Authorities (district and county as of April 2015), Metropolitan counties, Regions and Countries; Age data includes all ages, group breakdowns: 16-24, 16-17, 18-24, 25-49 and 50-64, individual ages: 18-22. Each year is also split by gender.  Generated on the 12th February 2019  This extends the data seen in the *Population Estimates* spreadsheet previously seen in the Workbooks. |
| 3 | *University of Wolverhampton Twitter feed (@wlv\_uni)*  *uow.json*  Generated on the 31st January 2019 |

You should carry out some investigations of two of the above datasets using the techniques used on the module to process and analyse the data. Lecture 2 outlined a Pipeline to follow, which you may wish to follow:

* Acquire data
* Prepare or Process data
* Analyse data
* Report or visualise data
* Act

You should write up the results of your investigation in a report of no more than 1500 words: This is a hard limit. This word count covers the main body of the report, including tables and figures. References and Bibliography would not be included in the word count.

The module has used the database systems Oracle, MongoDB and Hadoop; plus introduced the MapReduce framework. Bear in mind Buyya et al 2016 state in their book on Big Data that *“it should not be implied that NoSQL database are better database management systems than relational ones. The functionality and criteria to use each one is different”*. This statement could apply to Hadoop too.

Taking the above statement: “*The functionality and criteria to use each one is different”*, your report should include a matrix to compare the different database management systems (Oracle, MongoDB and Hadoop). Pick four key features you think are needed in a Big Data system and use the matrix to summarise where these three database systems are most relevant.

## The Report (70%)

Your report should cover the following areas:

* Introduction to Big Data
* Comparison of the main techniques used:
  + Oracle and Excel
  + MongoDB
  + MapReduce Framework
  + Hadoop

The comparisons should include examples from your investigation and the discussion should be in the context of handling big data.

* Matrix
  + Summarise 4 key features needed in Big Data and compare how the 3 database systems compare.
* Conclusion and way forward
* References and Bibliography

## Practical work (30%)

The practical work is to help you explore the three types of databases and evaluate where they are most appropriate. Document:

* What steps were taken to cleanse the data
* Any manipulation done to process the data before importing the data
* Analysis of the data to find results
* Presentation of the results

You must submit sufficient technical supporting material in an appendix to verify what you have done – this can be an appendix in a Word document, or a separate Notebook. You may use any analysis tools or techniques covered in the module.

Your submission must include your report and all supporting technical material required to replicate your findings. This could be in the form of a Word document which keeps a record what steps were taken to process and clean the data if necessary. If you are familiar with Python, you might want to consider using a Notebook to do this instead. You must provide sufficient information to allow others to follow and verify your findings.

Background reading is encouraged, particularly when introducing the topic of Big Data, or when discussing the tools and techniques. The University’s Harvard Referencing style should be used to reference any work.

See this website for details: <https://www.wlv.ac.uk/lib/skills-for-learning/referencing/>

## Hand-in date: 12th April 2019

A printed copy of the assessment should be submitted to the Student Office (one per team), **plus** each student should upload an electronic copy on Canvas. The version must be the same for each team.

**Ensure you keep a copy of your work.**

**Teamwork**

**PLEASE NOTE:**

You can work in groups of two for part 2 of the Portfolio, but no more than two. However, on no account should you work on the assignment with others groups to produce a larger group answer.

Ensure your coursework clearly states who the two team members are. Include a peer review that states what each member's contribution was to the assessment (should be equal). In the paper copy this statement should be signed by both students.

## Grade Attainment Criteria

See the 6CS030 marksheets for further details on the marking criteria for all of the above.

## Part 3 – 30 Marks

Time Constrained Assessment. This will be a test with 30 questions, which will be based on the Module’s quizzes.

Do note the questions will not be exactly the same as the quizzes.