

|  |
| --- |
| **COURSEWORK ASSESSMENT SPECIFICATION** |

|  |  |
| --- | --- |
| **Module Title:** | Big Data and Cloud Computing |
| **Module Number:** | KF7032 |
| **Module Tutor Name(s):** | Jeremy Ellman |
| **Academic Year:** | 2020-2021 |
| **% Weighting (to overall module):** | 100% |
| **Coursework Title:** | Big Data and Cloud Computing 2020-2021 Portfolio Assessment. |
| **Average Study Time** Required by a Student who hasfollowed **all** the module activities: | 50 study hours |

**Dates and Mechanisms for Assessment Submission and Feedback**

|  |
| --- |
| **Date of Handout to Students:**  Week Commencing 28th September 2020 |
| **Mechanism for Handout to Students:**  Blackboard. |
| Date and Time of Submission by Student:  7th January 2021 12:00 |
| **Mechanism for Submission of Work by Student:** *Blackboard*. |
| **Date by which Work, Feedback and Marks will be returned to Students:**  Within three working weeks after the submission date(s). |
| **Mechanism for return of assignment work, feedback and marks to students:**  Students will receive marks and feedback using gradebook on Blackboard after the marks have been moderated and approved by the module board. Individual feedback sessions will be available by prior appointment. |

The semi-formative elements of the portfolio constitute 30% of the assessment for this module and include, **group**, individual, and **peer assessed** work.

The practical element constitutes 70% of the assessment for this module and is an **individual** piece of work.

**Academic Integrity Statement: You must adhere to the university regulations on academic conduct. Formal inquiry proceedings will be instigated if there is any suspicion of plagiarism or any other form of misconduct in your work. Refer to the University’s Assessment Regulations for Northumbria Awards if you are unclear as to the meaning of these terms. The latest copy is available on the University website.**

* **Do NO**T submit code from other people or web sources as your own, this is **plagiarism**.
* Do NOT **buy** your assignments on the Internet or submit work written for you by others. This is **ghosting.**
* For the individual element Do NOT work with other students and submit identical code, this is **collusion**.
* Both plagiarism, ghosting, and collusion are academic misconduct, which is not allowed.

**Failure to submit:** The University requires all students to submit assessed coursework by the deadline stated in the assessment brief.  Where coursework is submitted without approval after the published hand-in deadline, penalties will be applied as defined in the University Policy on the Late Submission of Work. [https://www.northumbria.ac.uk/about-us/university-services/academic-registry/quality-and-teaching-excellence/assessment/](https://www.northumbria.ac.uk/static/5007/arpdf/lateappr)

## Aims

The aim of this assignment is to introduce a practical application of Big Data and Cloud Computing using a realistic big data problem. Students will implement a solution using an industry leading Cloud computing provider together with the distributed processing environment Apache Spark. This will involve the selection of problem appropriate Machine Learning algorithms and methods.

## Learning Outcomes Assessed:

### Knowledge & Understanding

1. Apply big data analytic algorithms, including those for visualization and cloud computing techniques to multi-terabyte datasets.
2. Critically assess data analytic and machine learning algorithms to identify those that satisfy given big data problem requirements

### Intellectual / Professional skills & abilities:

1. Critically evaluate and select appropriate big data analytic algorithms to solve a given problem, considering the processing time available and other aspects of the problem.
2. Design and develop advanced big data applications that integrate with third party cloud computing services

### Personal Values Attributes (Global / Cultural awareness, Ethics, Curiosity) (PVA):

1. Critically assess the relationship between knowledge and the ethical and social interpretation of primary research using big data.

## Definitions

**Portfolio Assignment**: A collection of pieces of work

**Individual Work**: Work carried out by one person only

**Group Work**: Work carried out collaboratively seeking to improve each other’s elements

**Peer Review**: Critical analysis and subsequent grading of a social equal’s work

**Semi-Formative**: Training tasks assigned course credit to reward and ensure engagement.

## Assignment Overview

The portfolio assignment is divided into components as follows:

|  |  |
| --- | --- |
| **Training Tasks** (30%) | Semi-formative elements of the portfolio constitute 30% of the assessment for this module and include, group, individual, and peer assessed work |
| **Combined Big Data Product and Report**: (70%) | Individual work – Combined Big Data Product and Report:  “A Critical Assessment of the Big Data Approach to Crime Analysis”. This activity assesses module learning outcomes 1, 2, 3, 4 & 5. This practical element is the final module assessment. |
|  |  |

**Training Tasks**

**Training Task 1: Peer Reviewed Task (24%)**

The objective of this task is to ensure that students have mastered these skills which are required for final module assessment:

1. Process a data set using the recommended software environment for the module.
2. Explaining the logical reasoning behind your code.

This work will be peer assessed as recommended the British Computer Society. That is, you will critically assess the work of fellow students (your peers) and THEY will assess yours.

In detail:

1. You will create a Jupyter notebook based the scenario below (which is derived from weekly worksheets 1-4) explaining your code using notebook embedded Markdown (i.e. formatted text, not just comments)
2. You will post your notebook to the module discussion board on Blackboard
3. You will then mark (i.e. peer review) the submission preceding yours on the discussion board, and the one following it, using the marking scheme below and post these mark sheets
4. Your mark for this task will be the average of your peer marks.

**Scenario:**

Suppose you are a police department with a limited budget. You plan to reduce road-traffic accidents by a one-month targeted advertising campaign.

Using the given dataset, which gender, age group, and month would be the largest target group as indicated by positive breath tests?

**Peer Reviewed Task Marking Scheme**

|  |  |
| --- | --- |
| The following marking scheme will be used for this assignment. A downloadable form will be placed on Blackboard that leaves space for comments. You should apply the general Marking Criteria to scale each element. This will be discussed in class. | |
| Description | Marks |
| **Peer Reviewed Task Marking Scheme:** | **(100)** |
| Introduction: The targeted advertising campaign task and approach taken to the problem | 10 |
| Component Selection and Data Pipeline Implementation (code) | 20 |
| Data Extraction and Filtering System running, test and diagnostics | 20 |
| Design, Development and reasoning behind analytic methods. | 20 |
| Detailed Analysis and consideration of the appropriateness of the solution for the initial problem | 10 |
| Evaluation and Conclusion | 10 |
| Style and Referencing | 10 |
|  |  |
| **Total Marks Available:** | **100** |

**Notes:**

1. Staff reserve the right to adjust marks that are disproportionately high or low, or where the two reviews differ by more than 20%.
2. Total marks from the two reviewers will be scaled to 24% of the module mark.

**Training Task 2: Group work participation Task (6%)**

The objective of this task is to derive background study materials for the big data product to be used by the whole class. These may include (but are limited to) reviewing the literature on crime and big data, examining published work on violent crime and its causes, technical approaches to crime and big data, relevant statistics and other computational methods. That is, to research the topic in general.

Working in teams of up to four students, each group will produce at least 2000 original words, plus 10 references to scientific conference or journal papers.

Since this is a group training task, your **participation** is assessed, rather than your content (Students will be able to receive staff feedback on content during taught sessions).

**Group work participation Task Marking Scheme**

Each group will score 6% of the module mark proportionally reduced by percentage of copied work as determined by Turnitin (threshold 10%), number of words less than 2000, number of references less than 10.

**Examples:**

Group A submit a total of 2100 words, plus 15 references which have a Turnitin similarity score of 8% (due to random matches). Each group member will score 6%.

Group B submit a total of 1500 words, plus 5 references which has a Turnitin similarity score of 20% (due to material copied from the Internet). Each group member will score:

(1500/2000) \* (110-20)/100\*(5/10) \*6% = 2%

**Big Data Product: Weapons and Drugs (Individual Work 70%)**

In the television documentary “Ross Kemp and the Armed Police” broadcast 6th September 2018 by ITV, multiple claims were made regarding violent crime in the UK.

These claims were:

1. Violent Crime is increasing
2. There are more firearms incidents per head in Birmingham than anywhere else in the UK
3. Crimes involving firearms are closely associated with drugs offences

In this assignment you will investigate these claims using real, publicly available data sets that will be made available to you and placed in Amazon S3. These include, but are not limited to:

1. **Street Level Crime Data** published by the UK Home Office. This dataset contains 19 million data rows giving a crime type, together with their location as a latitude and longitude.
2. **Land Registry Price Paid Data**: This gives the postcode of a property, the property type from an enumeration of D (Detached), S (Semi-Detached), T (Terraced), F (Flats/Maisonettes) and the price paid.
3. **Postcode Data**: This data set is based on material provided by the Ordinance Survey. It gives a latitude and longitude to every postcode. This is useful as it relates between the Land Registry Price Paid dataset postcode, and the original crime dataset

latitude/longitude.

#### Specifics

1. Process the data prepared for you using Apache Spark.
2. Filter the dataset so that crimes refer to relevant events only.
3. Using appropriate visualization methods, statistics, and machine learning, determine whether the claims made by Ross Kemp were true, false, or could not be determined.
4. Explain the reasoning behind your code so that it is clear what each block is intended to achieve, and why.
5. Report critically on the advantages, disadvantages, and limitations of the methods used.
6. Your submission will be a Jupyter Notebook containing both code (typically Python), and explanatory text (Markdown) limited to 2500 words (plus references). **References from scientific literature must be used** and your discussion **must be your own words. DO NOT CUT AND PASTE FROM THE INTERNET.**

.

### Feedback

Students will receive brief written feedback on the final submission together with the option to receive detailed verbal feedback on request.

## Hand-in Details

Submit your critical discussion using Turnitin assignment on Blackboard where further information will be given.

**Important:** Please do not include the original LARGE datasets in your submission. These are available to the module team already on Blackboard.

**Big Data Product Marking Scheme**

The following marking scheme will be used for this assignment.

|  |  |
| --- | --- |
|  | |
| Description | Marks |
| **Big Data Product Marking Scheme: Combined Code and Report** | **(70)** |
| Introduction: The Crime Analysis task and Approach taken to the problem | 10 |
| Component Selection and Data Pipeline Implementation (code) | 10 |
| Data Extraction and Filtering System running, test and diagnostics | 10 |
| Design, Development and reasoning behind use of **multiple** visualization methods, statistics, and machine learning Models | 10 |
| Selection, application, and reasoning behind use of statistical analysis and multiple evaluation measures | 10 |
| Detailed Analysis and consideration of the appropriateness of the solution for the initial problem | 10 |
| Evaluation and Conclusion | 5 |
| Style and Referencing | 5 |
|  |  |
| **Total Marks Available:** | **70** |
| . | |

## Marking Criteria

Since the elements above are wide ranging, general criteria are given that are applied as a percentage to each component of the portfolio. In the following, ‘writing’ is understood to apply both to coding and English.

|  |  |  |
| --- | --- | --- |
| **Percentage** | **General Criteria** |  |
| (0 – 29%) | A very poor contribution showing little awareness of subject area. Lack of clarity. Communication of knowledge is either inarticulate and or irrelevant.  Code fragments from the Internet may have replaced student written content to the extent that it is not possible to determine what the student has understood. Only partial functionality has been achieved. |  |
| (30 – 39%) | Knowledge is limited or superficial. Some awareness of concepts and critical appreciation are apparent, but there are major omissions or misunderstandings. Writing is not clear and there is no argument. Incorrect solutions or non-functioning software solutions have been given. |  |
| (40 – 49%) | Knowledge is barely adequate. Writing is fluent, but mostly, description and or assertion are used rather than argument or logical reasoning. A basic understanding of the key issues may have been demonstrated, but insufficient focus is evident in the work presented. Source code is functional, but poorly structured and commented. There may be some validation errors or security flaws. | Fail |
| (50 – 59%) | Knowledge base is up-to-date and relevant to an appropriate breadth and depth for level 7. The student has demonstrated the ability to apply theory and concepts, across domains and identify their interrelationship. A critical appreciation is demonstrated, which is supported by appropriate references. Writing is clear if a little uneven. Source code is functional, structured and commented. Code is valid and mostly secure. | Pass |
| (60 – 69%) | As above but there is clear evidence of independent thought and reasoned conclusions. Literature is fully supported by citation using appropriate references and there is development of a critical appreciation of opposing arguments. Presentation of work is fluent, focused and accurate. Source code is fully object‑oriented, secure, and completely-validates without being verbose. |  |
| (70 –100%) | Exceptional scholarship is demonstrated. There is a sustained ability to confront the current limits of knowledge in a relevant area or applied ‘real world’ contexts where demands of theory and practice may conflict. Argument is fluent, sustained, and convincing. Source code is of a professional standard. Clearly exceeds taught material. |  |