|  |  |
| --- | --- |
| **ASSESSMENT BRIEF** | |
| **Module Title:** | Big Data and Cloud Computing |
| **Module Code:** | PE7049 |
| **Academic Year / Semester:** | 2022-2023 / Semester 3 |
| **Module Tutor / Email (all queries):** | Baqar Rizvi / [baqar2.rizvi@northumbria.ac.uk](mailto:baqar2.rizvi@northumbria.ac.uk) |
| **% Weighting (to overall module):** | 100% |
| **Assessment Title:** | Big Data and Cloud Computing 2022-2023) Assessment. |
| **Date of Handout to Students:** | Monday 14th August 2023 |
| **Mechanism for Handout:** | Module Blackboard Site & Seminar in Week 6 |
| **Deadline for Attempt Submission by Students:** | Thursday 24th September 2023 23:59 |
| **Mechanism for Submission:** | Using Assignment submission on Blackboard. |
| **Submission Format / Word Count** | A Jupyter Notebook (.ipynb file) submission is required using Blackboard.  Also export your Notebooks as an html file and submit this to Turnitin. |
| **Date by which Work, Feedback and Marks will be returned:** | Within three working weeks after the submission date(s). |
| **Mechanism for return of Feedback and Marks:** | Mark and individual written feedback sheet will be uploaded to the Module Site on Blackboard. For further queries please email module tutor. |

The practical and combined report constitutes 100% 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 NOT submit code from other people or web sources as your own, this is

**plagiarism**.

* Do NOT work with other students and submit identical code, this is **collusion**.
* Do NOT buy your assignment on the Internet, or have your work done by someone
* else. This is **Ghosting**.
* Both Ghosting, plagiarism and collusion are academic misconduct, which is not

allowed and may result in you being asked to leave the University and lose any fees paid.

**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. Please ask student services (“Ask4Help”) for the latest version of this document. See [here.](https://www.northumbria.ac.uk/about-us/university-services/student-library-and-academic-services/quality-and-teaching-excellence/assessment/guidance-for-students/#:~:text=Late%20submission%20of%20work%20and%20extension%20requests&text=The%20student%20must%20submit%20an,the%20circumstances%2C%20must%20be%20provided.)

## 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 appropriate distributed processing environments such as Apache Spark. This will involve the provisioning and configuring of appropriate Cloud Computing resources and the selection of problem appropriate algorithms and visualization 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 and interpret primary research to identify its applicability to a given big data problem scenario.

## Assignment Overview

The assignment is divided into two components as follows:

|  |  |
| --- | --- |
| Big Data Product: practical and combined report (100%) | Individual work - Big Data Product practical and combined report: ‘Drugs and Guns’. This activity assesses all module learning outcomes. |
|  |  |

**Big Data Product: Weapons and Drugs (Individual Work 100%)**

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 that:

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

To solve this problem, you will use publicly available data sets that have been prepared for you and placed in online. 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. English Indices of Deprivation Data: The English Indices of Deprivation 2010 data set contains the rankings of measures of deprivation within small area level across England. The 32000 localities are ranked from the least to most deprived, scored on seven different dimensions of deprivation.

#### Specifics

1. Process the given data efficiently using Apache Spark on a cloud **Infrastructure as a Service (IaaS)** platform. A sample Jupyter Notebook has been provided on Blackboard.
2. Filter the dataset so that only relevant crimes are included.
3. Using appropriate techniques, determine whether Violent Crimes are increasing, decreasing, or are stable.
4. Determine whether there are more firearms incidents per head in Birmingham than anywhere else in the UK. Possession of firearms carries a mandatory prison sentence in the UK. Therefore, you may assume that a crime type of “Possession of weapons” whose outcome is “offender sent to prison” was a firearm incident.
5. Using appropriate techniques, determine whether firearms incidents are associated with drugs offences.
6. Select and prepare no more than four visualizations to support your analytic findings from (3).
7. Explain the reasoning behind your code so that it is clear what each block is intended to achieve (i.e., appropriately comment the command line).
8. Assess the three claims given and determine whether they are true, false, or cannot be determined.
9. Critically assess and report on the advantages, disadvantages, and limitations of the methods used.
10. Your submission will be a Jupyter Notebook containing both code (typically Python), and explanatory text (in Markdown format) limited to 2500 words (plus references). **References from the scientific literature must be used (please follow IEEE format)** 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 Jupyter Notebook using Blackboard. ALSO export the Notebook as html and submit this to Turnitin. Further information will be given on Blackboard.

**Big Data Product Marking Scheme**

The following marking scheme will be used for this assignment (Note that referred work is capped at 50%)

|  |  |
| --- | --- |
| Description | Marks |
| **Big Data Product Marking Scheme: Combined Code and Report** | **(100)** |
| 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 | 20 |
| Selection, application, and reasoning behind use of statistical analysis and multiple evaluation measures | 20 |
| Detailed Analysis and consideration of the appropriateness of the solution for the initial problem | 10 |
| Evaluation and Conclusion | 10 |
| Scientific References and Citation | 10 |
|  |  |
| **Total Marks Available:** | **100** |

## 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 is 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. |  |