**Processing Big Data**

**7CS516**



Dr. Oluwarotimi .W. Samuel

Contents

[Module Leader 3](#_Toc67579180)

[Key Dates and Details 3](#_Toc67579181)

[Description of the Assessment 3](#_Toc67579182)

[Assessment Content 4](#_Toc67579183)

[Assessment Rubric 4](#_Toc67579184)

[Anonymous Marking 6](#_Toc67579185)

[Assessment Regulations 6](#_Toc67579186)

[Formative 6](#_Toc67579187)

# Module Leader

Dr. Oluwarotimi .W. Samuel

Email: [o.samuel@derby.ac.uk](mailto:o.samuel@derby.ac.uk)

Office Hours: Monday 11am -12pm and Tuesday 11am -12pm

# Key Dates and Details

|  |  |
| --- | --- |
| **Assessment Type:** | Research and Data Analysis Implementation |
| **Assessment Weighting:** | 100% |
| **Word Count** | Up to 2500 ± 10% |
| **Learning Outcomes:** | 1 and 2 |
| **Submission Method:** | Blackboard/Turnitin |
| **Date Set:** | Uploaded in the BB module assessment space |
| **Submission Date:** | 12pm, end of week 13 at 19/01/2024 (Friday) |
| **Provisional Feedback Release Date:** | **In 14 days** |

# Description of the Assessment

This coursework would be completed in 2 parts. The first part consist of a structured report on systematic investigation and evaluation of notable current big data processing platforms and implementation of a data warehouse using one of these platforms. Besides, the second part of the assessment will focus on big data analytics tasks. That is, practical application of big data processing technologies and machine learning algorithm(s) implementation.

# Assessment Content

**Assessment Description:** The storage and processing requirement associated with Big Data have spurred various advancements in technologies and tools that could be leveraged to manage and process large-scale data such that business decisions, operations, and profits among others are optimized. At the core of these technologies is the cloud computing platforms, which Amazon Web Services (AWS), Microsoft Azure Platform (MAP), and Google Cloud Platform (GCP) are currently the leading providers in terms of market share. In view of the above, you are required to systematically investigate any two of the three cloud platforms with a case-study of Data Warehouse (DW) implementation. Afterward, you are required to report the DW’s schema and your evaluation results of the platforms investigated using no less than four core criteria that includes: (a) *Performance at Scale*, (b) *Elasticity* (c) *Ease of Use*, and (d) *Cost Efficiency*. Each of the four mentioned criteria is described in detail via this link (<https://www.firebolt.io/resources/cloud-data-warehouse-evaluation-guide>). Furthermore, at the end of your report, you are expected to recommend one cloud platform from the two you investigated and state your reason (s) for choosing such a platform for Processing and Managing Big Data. Upon investigating the characteristics of the cloud platforms, you should be able to process Big Data with the goal of performing analysis using Python that could reveal insights capable of driving efficient business decisions. The dataset and Python-SQL driven DW implementation will be explained in the practical/tutorial session.

**Deliverables:**

Report on evaluation of big data cloud platform and data warehouse demo implementation process. The report should be structured as follows:

1. **Introduction 10%**

This section of the report should primarily focus on the purpose and scope of the report, and should be documented as follows:

* Brief description of the cloud platforms investigated and evaluated
* The criteria employed for the evaluation
* The investigation methodology

1. **Platform investigation 40%**

* Detailed description of the evaluation on each platform according to the defined criteria should be documented in the report.
* The comparative results obtained should be documented

1. **Big Data processing and analytics 40%**

* Performing big data processing tasks by using NoSQL, PySpark or similar techniques to specifically carry out data analysis via one of the cloud platforms or conduct a simulation on your own personal computer. The analysis should mainly focus on data pre-processing, exploratory data analysis (EDA), building a machine learning model (s) for classification/prediction task (s).
* Provide explanations and screenshots of the implementation to support this section. Critical discussion on the reason special algorithms are selected for the implementation should be discussed.
* The dataset will be provided. The dataset is relatively big for assessment purpose, and it can be downloaded from the Assessment section of the module on blackboard.

1. **Conclusion 10%**

* Summarisation
* Experience (what you have learnt from the assessment) discussion
* Future work (what can be improved)

Submission must be in Word (.doc, .docx) or PDF file format.

All these will be reviewed, assessed and feedback provided on the blackboard.

# Assessment Rubric

The assessment rubric on the next page shows the complete criteria of the CW and how you will be assessed. We will explore the content of the rubric together in a synchronous session on first week session.

When the assessment is returned, you will receive a digital version of the rubric showing how you performed against each criterion. You will also receive a short individual audio file that highlights both the strengths of the work and your key areas for development. There will be an opportunity to discuss these with your personal academic tutor.

**Rubric**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 20-39 | 40-49 | 50-59 (Passing grade band) | 60-69 | 70-100 |
| **Introduction (10)** | Not thought out or planned. Impossible to see where this is going | Platforms are selected and criteria are defined, but with very little discussions on the selection process on criteria and platforms. Methodology is not defined for investigation. | Platforms are selected and criteria are defined with some discussions on the selection process on criteria and platforms. The methodology is defined but not explained well. | Good justification and discussions on the platforms & criteria for their selections. Well defined methodology for the investigation. | Great justification and in-depth discussions on the platforms & criteria for the selections. Great defined methodology for investigation task. (70-80)  + self-defined & suitable criteria (+10)  + The methodology is fantastic and explained well (+10) |
| **Platform investigation (40)** | No evidence of any self-led investigation on any suitable big data platform | Investigations are difficult to understand and no clear criteria and methodology.  Or less than 2 platforms were investigated. | No less than 2 platforms were investigated with satisfactory discussions on the defined criteria. But overall, the discussions can be improved on self-thinking, criterial analysis and in-depth discussions. | No less than 2 platforms were investigated with satisfactory discussions on the defined criteria. The discussions include self-thinking, criterial analysis, and in-depth. | No less than 2 platforms were investigated with satisfactory discussions on the defined criteria. (70-80)  +Strong self-thinking and critical analysis (+10)  +final evaluation result is valuable to guide other people to do the work (+10) |
| **Big data analysis implementation (40)** | No evidence of any sufficient implementation on all the tasks | Not completed the three tasks in any level but one (40-45) or two (45-49) tasks were implemented but contains lots of issues. | Two (45-49) tasks were implemented but contains lots of issues with **PySpark**.  Or  At least two tasks were implemented at a good level and runnable to show the reasonable results. But no critical analysis (e.g. comparing to alternative methods or discussions of selections) | At least two tasks were implemented at a good level and runnable to show the reasonable results using **PySpark**. But no critical analysis (e.g. comparing to alternative methods or discussions of selections) (60-65)  Or All three tasks were implemented at a good level and runnable to show the good results with good critical analysis. (60-65)  Or All three tasks were implemented using **PySpark** at a good level and runnable to show the good results with good critical analysis. (65-69) | All three tasks were implemented at a good level and runnable to show the high-performance results with great critical analysis. (70-80)  + **PySpark** implementation (+10)  + alternative methods are implemented in critical analysis (+10) |
| **Conclusion (10)** | Not having a good summary at the end, no experience and future work discussions. | The summary is too basic | having a good summarisation at the end with little experience and future work discussions. | having a good summarisation at the end with good experience and future work discussions. | having a good summarisation at the end with fantastic experience and future work discussions. (70-80)  + some valuable outcome and recommendations (+10)  + future work including extensions on any other criteria evaluation (+10) |

# Anonymous Marking

You must submit your work using your **student number** to identify yourself, not your name. You must not use your name in the text of the work at any point. When you submit your work in Turnitin you must submit your student number within the assignment document and in the *Submission title* field in Turnitin. A video showing how to do this can be found here (link).

# Assessment Regulations

The [University’s regulations, policies and procedures](https://www.derby.ac.uk/about/academic-regulations/) for students define the framework within which teaching, and assessment are conducted. Please make sure you are familiar with these regulations, policies and procedures.