

School of Design and Informatics

**CMP406 – Enterprise Systems Engineering**

**Module Tutor: Mr. David McLuskie**

Semester 1 - Resit

***Instructions:***

Please see below

**Submission**:

**UNIT 1:**

The portfolio should be posted electronically (uploaded) **as ONE pdf file** on the module MyLearningSpace site CMP406 – Enterprise Systems Engineering and Big Data no later than:

**Thursday 4th April 2024, 12 Noon**

**Feedback for Unit 1 will be issued 15 working days after submission**

**Late submission:** Work submitted up to 5 days following the submission deadline will be accepted in the absence of valid cause but will be subject to a penalty of 1 grade (-0.5 GPA) for each day it is late. Work submitted after 5 days without valid cause will be treated as a non-submission and will be awarded a NS grade.

## Learning Outcomes

This coursework covers the following learning outcomes for the module:

1. Explain and exemplify the enterprise and information architectures of an enterprise system, including the hardware and software requirements.   
4. Describe and exemplify the methods and methodologies appropriate to systems security in an enterprise system.

## Generative AI

Use of Generative AI Tools such as ChatGPT, DALL-E, Bard etc. is **explicitly prohibited** for Assessment Unit 1 in CMP406. The output gleaned or generated from the tools mentioned and others that are similar relates to sources already published/available. Consequently, using these tools without correct referencing is deemed plagiarism. Also, be aware that the information obtained can be inaccurate or incomplete. Thus, all submitted work should be your own. If the assessment is found to have been plagiarised or to have used unauthorised AI tools, you will be referred to the Student Disciplinary Officer within the School and this may result in an Academic Misconduct charge.

You may use specific tools, such as Grammarly for aiding Grammar and Structure.

* It is recommended that you keep a record of your interaction with these tools.
* It is recommended that you have evidence of drafts of your work.

CMP406: Enterprise System Engineering and Big Data October 2023

Task 1

# General Instructions

For your submission you should upload a PDF document which includes a list of your answers as well as the code statements that you used to generate your answers. You should also upload a video where you will explain how your code works and show the execution of your queries along with the results.

The coursework needs to be completed by Thursday 4th April 2024, 12 noon. Your answers and code should be uploaded as a PDF document to MyLearningSpace. Your video should be uploaded to MyLearningSpace as part of your submission.

Task 1 is worth 15% of your overall module grade for CMP406 and forms the first part of your lab portfolio.

# Assignment

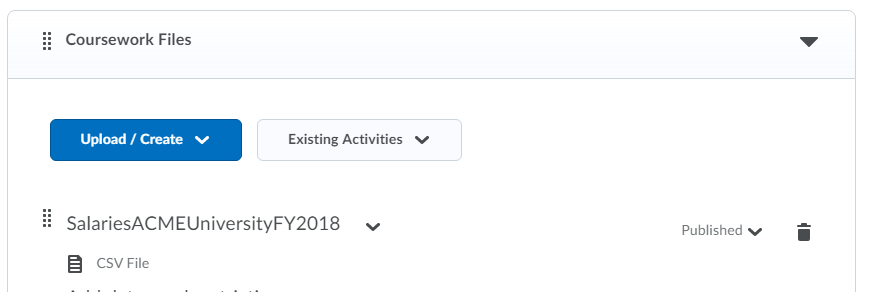
ACME University has suspected that somebody has artificially inflated some of the employees pay but considering the size of the university and the number of people that are employed they are finding it difficult to process all of the data to find the culprits. They have hired you to perform a forensic investigation of the universities staff payroll and they want you to identify potential suspects who are being paid significantly more than what their role is worth.

You have identified that the processing and storage of the data can be accomplished by using Hadoop. Your task is to choose a programming language from the Hadoop ecosystem, but not spark, and develop a program or a series of programs which can be executed on Hadoop to identify the suspects who are being paid more than what their grade is worth.

In addition to the above task, ACME University would also like you to use your chosen programming language to produce the following statistics about the staff that are employed by the university:

* The number of staff that are working full and part-time.
* The number of staff that are working in each service type.
* The average salary in each Service type.
* The number of staff that are in each job role.
* Determine how many staff are earning below the average salary for the university.

The data file that ACME would like you to use is called **SalariesACMEUniversityFY2018** andit is located under the week 5 folder on MyLearningSpace.



**Grading sheet for CMP406 – Task 1**

Student Name:

| **Grade/mark**  **Criteria** | **A**  **4.5-4.0** | **B**  **3.5-3.0** | **C**  **2.5-2.0** | **D**  **1.5-1.0** | **MF**  **0.5** | **F**  **0.0** | **NS** | **Comments** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Explanation**  Able to give a clear explanation as to what is happening with the code.  Code is appropriately structured and commented |  |  |  |  |  |  |  |  |
| **Demonstration**  Able to demonstrate a working application or series of applications which can be run on Hadoop |  |  |  |  |  |  |  |  |
| Overall comments  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  **GRADE:** |

| **Literal grade** | **Grade Point** | **Evaluative descriptor** | **ECTS** |
| --- | --- | --- | --- |
| A+ | 4.5 | Excellent overall.   * Demonstrates an excellent grasp of the subject matter. * Excellent capacity for original and creative enquiry. * Excellent ability to critically evaluate, analyse, synthesise and integrate complex information. * Excellent communication skills.   Exceptional in at least one of the above. | **A** |
| A | 4 | Excellent overall.   * Demonstrates an excellent grasp of the subject matter. * Excellent capacity for original and creative enquiry. * Excellent ability to critically evaluate, analyse, synthesise and integrate complex information. * Excellent communication skills. | **A** |
| B+ | 3.5 | Very good overall.   * Demonstrates a very good grasp of the subject matter. * Very good capacity for original and creative enquiry. * Very good ability to critically evaluate, analyse, synthesise and integrate complex information. * Very good communication skills.   Excellent in at least one of the above but overall performance deemed to be very good. | **B** |
| B | 3 | Very good overall.   * Demonstrates a very good grasp of the subject matter. * Very good capacity for original and creative enquiry. * Very good ability to critically evaluate, analyse, synthesise and integrate complex information. * Very good communication skills. | **B** |
| C+ | 2.5 | Good overall.   * Demonstrates a good grasp of the subject matter. * Good capacity for original and creative enquiry. * Good ability to critically evaluate, analyse, synthesise and integrate complex information. * Good communication skills   Very good in at least one of the above but overall performance deemed to be good. | **C** |
| C | 2 | Good overall.   * Demonstrates a good grasp of the subject matter. * Good capacity for original and creative enquiry. * Good ability to critically evaluate, analyse, synthesise and integrate complex information. * Good communication skills | **C** |
| D+ | 1.5 | Satisfactory overall.   * Demonstrates a satisfactory grasp of the subject matter but limited grasp in some areas * Satisfactory capacity for original and creative enquiry. * Satisfactory ability to critically evaluate, analyse, synthesise and integrate information. * Satisfactory communication skills | **D** |
| D | 1 | Adequate.  Achievement of all threshold standards but grasp of some subject areas and graduate attribute development may be more limited. | **D** |
| MF | 0.5 | Marginal fail.  Performance just below the threshold standard. A reasonable expectation that a pass is achievable by reassessment without the need to repeat the module. | **MF** |
| F | 0.0 | Fail. Performance well below the threshold level. Some limited evidence of achievement of the outcomes. | **F** |
| NS |  | No assessments submitted. | **NS** |

CMP406: Enterprise System Engineering and Big Data October 2023

Task 2

# General Instructions

This task is comprised of a small project to create a small cluster which is running a Spark application to perform some rudimentary language detection on a dataset. You are also required to secure the cluster to restrict access to the datasets stored on HDFS. As part of this task, you should keep a log book to outline what you have done and especially any problems you many have encountered. There is no minimum or maximum length of log book and it can contain many screenshots etc.

The purpose of the log is to show that you completed the work yourself and have an understanding of core concepts. It will contribute to your grade for each area of the marking scheme, so you should focus on detailing any areas which you may have struggled with.

The coursework needs to be completed by Thursday 4th April 2024, 12 noon Your log book should be uploaded as a PDF document to MyLearningSpace. You should also upload a video where you will demonstrate your coursework. The full requirements are listed on the next page. The video should focus on the following areas:

1. Demonstrate your cluster working (including showing the relevant web interfaces) and explain how you went about setting the cluster. You must show the following:
   1. HDFS running – show the web interface status page
   2. YARN running – show the web interface status page
   3. Run the Hadoop Pi example application, and show that this is running, and then show that it has finished on the YARN status page.
   4. Demonstrate creating a file and copying it to your HDFS system.
2. Explain in detail how your Spark code works and demonstrate your code working
3. Demonstrate Kerberos authentication working on your cluster and show that the Pi program can be executed successfully on it. You should also explain the steps that you took to get Kerberos working, and this should be backed up with examples of the configuration changes made. Use the HDFS status page to show that the Security status is “on”.

Your video should be uploaded to MyLearningSpace as part of your submission.

Task 2 is worth 35% of your overall module grade for CMP406 and forms the second part of your lab portfolio.

# Task 2 Requirements Breakdown

## Spark Application

For the first part of this task, you are required to write a short Spark application which will analyse a dataset and will perform some rudimentary language identification.

The dataset for the application will come from Project Gutenberg which provides a large number of free ebooks. ([www.gutenberg.org](http://www.gutenberg.org)). From this site you should download a selection of books in English, French, and German and store them on HDFS.

Your Spark application should be written to look at the books in a directory and give an indication of what language is most common in the directory (English, French or German). For the purpose of this task, this need only be very basic. You may want to look at conducting a basic word count on the books and checking which words appear very frequently across the different languages.

You should write the Spark application using the Scala programming language. Java and Python can be used but you **will not** achieve full marks for using these languages.

You should compile your Spark to a JAR file and be able to demonstrate that you can launch your application on YARN using spark-submit.

Please include your Spark source code in your log book.

## Deploying Your Application

For the second part of this task, you should set up a small (3 nodes) cluster which is running Spark, YARN, and HDFS. You must try and set this up from scratch by configuring all parts of the system. Do not use the Cloudera VM you were provided with for this task, as the RAM requirement makes this difficult.

You should take note of the steps you take in doing this in your log book. This will act as part of your submission to MyLearningSpace. You should detail any problems you had, as this can reinforce your grade should your cluster fail to work and will allow for more detailed feedback.

## Securing the Cluster

Once you have a small cluster running, you should set up Kerberos authentication to secure your HDFS datasets. You should be able to demonstrate that access to your datasets is restricted to certain users.

**Grading sheet for CMP406 – Task 2**

Student Name:

| **Grade/mark**  **Criteria** | **A**  **4.5-4.0** | **B**  **3.5-3.0** | **C**  **2.5-2.0** | **D**  **1.5-1.0** | **MF**  **0.5** | **F**  **0.0** | **NS** | **Comments** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Completed Spark language detection application [40%]**  Able to demonstrate a working application which performs basic language detection |  |  |  |  |  |  |  |  |
| **Cluster setup and Spark application submission [40%]**  Able to demonstrate a compiled Spark application can be submitted to a cluster which has been set up, and is using YARN. |  |  |  |  |  |  |  |  |
| **HDFS secured using Kerberos authentication [20%]**  Can demonstrate that Kerberos authentication has been implemented on their cluster to secure access to HDFS. |  |  |  |  |  |  |  |  |
| Overall comments  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  **GRADE:** |

| **Literal grade** | **Grade Point** | **Evaluative descriptor** | **ECTS** |
| --- | --- | --- | --- |
| A+ | 4.5 | Excellent overall.  Spark application has been written well and is explained by the student clearly.  Student had no problems running the application. Code written in Scala.  Cluster is fully operational, and the Spark application runs on their cluster. Student is confident when explaining their cluster and can clearly answer any questions.  Kerberos is fully operational and cluster works correctly. Student shows a clear understanding.  Exceptional in some aspects. | **A** |
| A | 4 | Excellent overall.  Spark application has been written well and is explained by the student clearly.  Student had no problems running the application. Code written in Scala.  Cluster is fully operational, and the Spark application runs on their cluster. Student is confident when explaining their cluster and can clearly answer any questions.  Kerberos is fully operational and cluster works correctly. Student shows a clear understanding. | **A** |
| B+ | 3.5 | Very good overall.  Spark application works and can be mostly explained by the student. Student may have struggled a little to launch application or explain code. Code written in Scala.  Cluster is fully operational but may be missing Spark. YARN and HDFS functionality shown to work correctly. Student demonstrates functionality well.  Kerberos is operational on HDFS but the student cannot get apps to run correctly. Demonstrates a good understanding of Kerberos concepts.  Excellent in some aspects but overall performance deemed to be very good. | **B** |
| B | 3 | Very good overall.  Spark application works and can be mostly explained by the student. Student may have struggled a little to launch application or explain code. Code written in Scala.  Cluster is fully operational but may be missing Spark. YARN and HDFS functionality shown to work correctly. Student demonstrates functionality well.  Kerberos is operational on HDFS but the student cannot get apps to run correctly. Demonstrates a good understanding of Kerberos concepts. | **B** |
| C+ | 2.5 | Good overall.  Spark application does not work fully, however the code provided shows a good framework, and the student has explained the code with a clear understanding of a potential solution.  Code is still packaged as a JAR (unless Python used) and should run partially. Code written in Scala, Python or Java.  Cluster is only partially operational. Either HDFS or YARN is running, as evidenced by the UI web page showing the correct information. Student can explain the problems they had in setting up. This should be reinforced by log book evidence.  Kerberos is not operational on HDFS (security is off, or no HDFS UI) but the student has set up the Kerberos server correctly and can demonstrate adding new principals and using ‘kinit’ to get new tickets. Reinforced by log book.  Very good in some aspects but overall performance deemed to be good. | **C** |
| C | 2 | Good overall.  Spark application does not work fully, however the code provided shows a good framework, and the student has explained the code with a clear understanding of a potential solution.  Code is still packaged as a JAR (unless Python used) and should run partially. Code written in Scala, Python or Java.  Cluster is only partially operational. Either HDFS or YARN is running, as evidenced by the UI web page showing the correct information. Student can explain the problems they had in setting up. This should be reinforced by log book evidence.  Kerberos is not operational on HDFS (security is off, or no HDFS UI) but the student has set up the Kerberos server correctly and can demonstrate adding new principals and using ‘kinit’ to get new tickets. Reinforced by log book. | **C** |
| D+ | 1.5 | Satisfactory overall.  The application does not work fully and the solution proposed by the student may be incomplete. Code may not be packaged as a JAR. Student has some understanding of how to complete a solution. Code written in Scala, Python or Java.  The student has struggled to get any part of the cluster working correctly, but at a minimum, has managed to get YARN or HDFS working, but the worker nodes may not be communicating correctly. Basic understanding and a reasonable attempt to set up the cluster should be evidenced by the log book.  Kerberos is installed and has been set up on a very basic level. The student can evidence this using administration tools. They have a basic understanding of how Kerberos works. Reinforced by log book.  Some elements are good. | **D** |
| D | 1 | Adequate.  The application does not work fully and the solution proposed by the student is incomplete. Code may not be packaged as a JAR. Student has some understanding of how to complete a solution. Code written in Scala, Python or Java.  The student has struggled to get any part of the cluster working correctly, but at a minimum, has managed to get YARN or HDFS working, but the worker nodes may not be communicating correctly. Basic understanding and a reasonable attempt to set up the cluster should be evidenced by the log book.  Kerberos is installed and has been set up on a very basic level. The student can evidence this using administration tools. They have a basic understanding of how Kerberos works. Reinforced by log book.  Achievement of all threshold standards but grasp of some subject areas and graduate attribute development may be more limited. | **D** |
| MF | 0.5 | Marginal fail.  Performance just below the threshold standard. A reasonable expectation that a pass is achievable by reassessment without the need to repeat the module. | **MF** |
| F | 0.0 | Fail. Performance well below the threshold level. Some limited evidence of achievement of the outcomes. | **F** |
| NS |  | No assessments submitted. | **NS** |