Assessment Brief

Academic Year	**2022/2023**
Semester	**1**
Module Number	**CMM-705**
Module Title	**Big Data Programming**
Assessment Method	
Deadline (time and date)	**24th December 2022. 1.30 PM UK Time**
Submission	**Assessment Dropbox in the Module Study Area in CampusMoodle.**
Word Limit **(see Assessment Word Limit Statement)**	**No word limit**
Module Co-ordinator	**Sajith Ravindra**

What knowledge and/or skills will I develop by undertaking the assessment?

This coursework examines students' ability to design, implement, and deploy a big data analytics solution. This exercise encourages students to engage in activities such as designing the solution architecture, selecting the required technologies/products to realize the solution architecture and achieve systems functional and non-functional goals, implementing and deploying the solution, and presenting big data analytics results.

On successful completion of the assessment students will be able to achieve the following Learning Outcomes:

1. Critically assess a big data problem and provide a robust solution
2. Employee batch analytics techniques effectively
3. Using existing machine learning algorithms to solve a problem
4. Being able to present data/analysis findings using appropriate mean

Please also refer to the Module Descriptor, available from the module Moodle study area.

What is expected of me in this assessment?

Dataset

This dataset includes ball-by-ball data records of cricket matches that were played in the Indian Premier League(IPL) from the year 2008 through 2020. You can find the complete dataset in the ipl-data.zip

Columns details

Column Name	Description
id	A unique identifier given to each match
inning	Indicates the innings of the match
over	The over to which the record belongs to
ball	The ball of the over
batsman	The batsman who faced the ball at the striker's end
non_striker	The other batsman at the non-strikers end
bowler	The bowler who delivers the ball
batsman_runs	Runs the batsman has scored
extra_runs	Extra runs given a way
total_runs	Total runs gained for the delivery
is_wicket	Indicate if a wicket was taken in the ball
dismissal_kind	How the wicket was taken
player_dismissed	The batsman who’s wicket was taken
fielder	The fielder who contributed to taking the wicket
extra_type	The reason for extra runs
batting_team	The name of the batting team
bowling_team	The name of the bowling team

Questions

Designing a Solution Architecture to analyze the data on Cricket Matches. Assume that the systems receive data in real time. Therefore, your system design should be able to perform historical and real-time data analysis.

What is expected of me in this assessment?

[Total Marks 25]

1. System diagram with proposed big data tools, to collect data from the systems in real-time, store them in scalable storage, and process periodically to produce summarised information to visualize in a dashboard.

   [Marks 15]

2. Describe the role of each component and how the overall architecture

   [Marks 10]

2. Data Analysis ***[Total Marks 40]***

In this section, you are developing an analytics app for cricket analysis and team management groups to understand how players and teams have performed. This analysis provides high-level summarizations and insights allowing the sports analyst and the authorities to understand how teams and individual players have performed and identify areas of improvement for each team and individual player

2.1. Analyze the following using Hadoop MapReduce *[Marks 10 (5 for each)]*

1. The number of deliveries in which a wicket was taken, extra runs were given and no runs were scored.
2. The Total number of wickets taken by each team

2.2. Analyze the following using Hive or Pig *[Marks 15 (7.5 for each)]*

1. Top 10 teams, based on the total number of runs scored (the team which has scored the most runs is the best team)
2. The average runs scored in each over of an innings (i.e., the overall average across the dataset for each over from over 1 to 20).

2.3. Analyze the following using Spark *[Marks 15 (7.5 for each)]*

1. Percentage of players who have scored 50 runs or more in a single inning.
2. The total number of matches won, lost, and drawn by each team.

3. Performing Machine Learning model using Spark MLlib ***[Total Marks 19]***

Build a model that predicts the average runs expected to be scored in the first 6 overs if TeamA team play against TeamB, where TeamA is the batting team and TeamB is the bowling team

What is expected of me in this assessment?

Example: *Predict what's the expected score in the first 6 overs if Mumbai Indians bat against Kolkota Night Riders.*

Use 80% of the data for training and 20% of the data for validation. Clearly state the steps you've followed for your analysis.

4. Presentation of the analysis ***[Total Marks 16]***

A static web page or a presentation of data using some other visualization tool to view the results below mentioned results gathered from your analysis

[Marks 4 for each]

1. The number of deliveries in which a wicket was taken, extra runs were given and no runs were scored.
2. Top 10 teams, based on the total number of runs scored (the team which has scored the most runs is the best team)
3. Percentage of players who have scored 50 runs or more in a single inning.
4. The total number of matches won, lost, and drawn by each team.

This can be a static web page/dashboard with hard-coded values obtained from your analysis, it's *NOT* required to dynamically fetch data and display.

The main assessment criterion is how well the data is presented in an understandable manner. Also address UX, and UI aspects in your design when presenting the data.

Task(s) - format

You will be required to submit the following two deliverables to Campus Moodle.

- A report compiled including the details mentioned above and with a properly completed cover sheet should be submitted in ***PDF format*** (please DON'T zip your report or upload any other format).
- A .zip archive covering the ***source code*** of all your implementations, including your
  - Java files (Map Reduce)
  - Text file (Hive/Pig)
  - Zepplin/Jupyter/etc. Notebooks or scripts (Spark and ML)
  - HTML, CSS, and JavaScript files. (Dashboard)

What is expected of me in this assessment?

This should be submitted to ***.zip format*** (please DON'T use other archive formats). Don't include the data set, built jar files, or any other artifacts 3rd party artifacts.

Report Format

Your report should include,

- For part 1:
  - Deployment architecture to collect, analyze the data and present the results to end-users (show the software/tools that can be used for implementing the solution)
  - And reasoning on the proposed architecture and why each technology tool was selected
- For part 2, for ***each question***:
  - The final result/output. In case the result consists of many rows add only the topmost part of the output.
  - And the code listings of the implementation.
- For part 3:
  - Step-by-step breakdown of the steps followed.
  - For each step include code listings and screenshots.
- For part 4:
  - Screenshots of the dashboard

How will I be graded?

A grade will be provided for each criterion on the feedback grid which is specific to the assessment.

The overall grade for the assessment will be calculated as follows:

Final Marks = (Marks given for the report and submissions * 0.8) + (Marks for viva * 0.25)

A	Final Mark >= 80
B	80 > Final Mark >= 60
C	60 > Final Mark >= 50

How will I be graded?	
D	50 > Final Mark >= 40
E	40 > Final Mark >= 30
F	30 > Final Mark
NS	Non-submission.

Feedback grid

GRADE	A	B	C	D	E	F
DEFINITION / CRITERIA (WEIGHTING)	**EXCELLENT** Outstanding Performance	**COMMENDABLE/VERY GOOD** Meritorious Performance	**GOOD** Highly Competent Performance	**SATISFACTORY** Competent Performance	**BORDERLINE FAIL**	**UNSATISFACTORY** Fail
CRITERION 1 **(x %)** **Grade:** □						
CRITERION 2 **(x %)** **Grade:** □						
CRITERION 3 **(x %)** **Grade:** □						
CRITERION 4 **(x %)** **Grade:** □						

Coursework received late, without valid reason, will be regarded as a non-submission (NS) and one of your assessment opportunities will be lost.

What else is important to my assessment?

What is plagiarism?

“Plagiarism is the practice of presenting the thoughts, writings or other output of another or others as original, without acknowledgement of their source(s) at the point of their use in the student’s work. All materials including text, data, diagrams or other illustrations used to support a piece of work, whether from a printed publication or from electronic media, should be appropriately identified and referenced and should not normally be copied directly unless as an acknowledged quotation. Text, opinions or ideas translated into the words of the individual student should in all cases acknowledge the original source” (RGU 2022).

What is collusion?

“Collusion is defined as two or more people working together with the intention of deceiving another. Within the academic environment this can occur when students work with others on an assignment, or part of an assignment, that is intended to be completed separately“ (RGU 2022).

For further information please see Academic Integrity.

What is the Assessment Word Limit Statement?

It is important that you adhere to the Word Limit specified above. The Assessment Word Limit Statement lists what is included and excluded from the word count, along with the penalty for exceeding the upper limit.

What if I’m unable to submit?

- The University operates a Fit to Sit Policy which means that if you undertake an assessment then you are declaring yourself well enough to do so.
- If you require an extension, you should complete and submit a Coursework Extension Form. This form is available on the RGU Student and Applicant Forms page.
- Further support is available from your Course Leader.

What else is important to my assessment?

What additional support is available?

- RGU Study Skills provide advice and guidance on academic writing, study skills, maths and statistics, and basic IT.
- RGU Library guidance on referencing and citing.
- The Inclusion Centre: Disability & Dyslexia.
- Your Module Coordinator, Course Leader, and designated Personal Tutor can also provide support.

What are the University rules on assessment?

The University Regulation ‘A4: Assessment and Recommendations of Assessment Boards‘ sets out important information about the assessment and how it is conducted across the University.