

Assessment Brief, 2022-23

1. An overview of the assignment:

1.1. General Information

Module title	Big Data Tools & Techniques
CRN	41141 (September) 50194 (January)
Level	7
Assessment title	Insert details
Weighting within module	This assessment is worth 100% of the overall module mark.
Assessment set by (Family name Alphabetical order)	Kaveh Kiani Taha Mansouri Nathan Topping
Submission deadline date and time	28th/April/2023 4 pm For coursework assessments only: students with a Reasonable Adjustment Plan (RAP) or Carer Support Plan should check your plan to see if an extension to this submission date has been agreed.

1.2. Learning outcomes of this assessment

The learning outcomes covered by this assignment are:

- Provide a broad overview of the general field of 'big data systems'
- Developing specialised knowledge in areas that demonstrate the interaction and synergy between ongoing research and practical deployment of this field of study.

1.3. Key skills to be assessed

This assignment aims at assessing your skills in:

- The usage of common big data tools and techniques
- Your ability to implement a standard data analysis process
 1. Loading the data
 2. Cleansing the data
 3. Analysis
 4. Visualisation / Reporting
- Use of Python, SQL, and Linux terminal commands

1.4. Recommended Reading

The module notes complimented by tools and techniques covered in other modules are sufficient literature for completing this assignment successfully.

1.5. Equipment and Facilities to be Used

For this assignment, only the Databricks platform are to be used. All processing must be done via executable notebooks, scripts and code, and these must be stored and included with the submission.

1.6. Workload

For the successful completion of this assignment, a total of 120 hours should be budgeted.

1.7. General instructions

You will follow a typical data analysis process:

1. Load / ingest the data to be analysed

2. Prepare / clean the data
3. Analyse the data
4. Visualise results / generate report

For all steps you will use environments that have been used within this module.

1.8. General notes

- The assignment must be completed on your own: this includes all the programs and report. By the act of following these instructions and handing your work in, it is deemed that you have read and understand the rules on plagiarism as written in the academic handbook (see also the Unfair means point below).
- The assignment must be completed on time. If you submit work late, it will be marked according to the University's late submission policy.

1.9. Assessment Criteria

Information on the assessment criteria of this assignment is provided in the rubric at the end of this document.

1.10. Word count of the report

Your assessment should be no more than 5,000 words in total. The report for Task 1 should be no more than 3,000 words and the report for Task 2 should be no more than 2,000 words.

1.11. Academic Integrity and Referencing

Students are expected to learn and demonstrate skills associated with good academic conduct (academic integrity). Good academic conduct includes the use of clear and correct referencing of source materials. Here is a link to where you can find out more about the skills which students need:

[Academic integrity & referencing](#)

[Referencing](#)

Academic Misconduct is an action which may give you an unfair advantage in your academic work. This includes plagiarism, asking someone else to write your assessment for you or taking notes into an exam. The University takes all forms of academic misconduct seriously.

1.12. Assessment Information and Support

Support for this Assessment

You can obtain support for this assessment by contacting the module team via email on:

N.J.Topping@salford.ac.uk or t.mansouri@salford.ac.uk or k.kiani@salford.ac.uk

You can find more information about understanding your assessment brief and assessment tips for success [here](#).

Assessment Rules and Processes

You can find information about assessment rules and processes in Blackboard in the [Assessment Support](#) module.

Develop your Academic and Digital Skills

Find resources to help you develop your skills [here](#).

Concerns about Studies or Progress

If you have any concerns about your studies, contact your Academic Progress Review Tutor/Personal Tutor or your Student Progression Administrator (SPA).

askUs Services

The University offers a range of support services for students through [askUS](#) including Disability and Learner Support, Wellbeing and Counselling Services.

Personal Mitigating Circumstances (PMCs)

If personal mitigating circumstances (e.g. illness or other personal circumstances) may have affected your ability to complete this assessment, you can find more information about the Personal Mitigating Circumstances Procedure [here](#). Independent advice is available from the Students' Union Advice Centre about this process. Click [here](#) for an appointment to speak to an adviser or email advicecentre-ussu@salford.ac.uk.

1.13. Reassessment

If you fail your assessment and are eligible for reassessment you can do one reassessment attempt. You will be given the same assignment and enough time to do the assignment again, but your mark will be capped to 50.

For students with accepted personal mitigating circumstances for absence/non submission, this will be your replacement assessment attempt (but your mark will not be capped).

We know that having to undergo a reassessment can be challenging however support is available. Have a look at all the sources of support outlined earlier in this brief and refer to the [Personal Effectiveness](#) resources.

2. Tasks

You will be given 2 tasks with separate datasets a set of problem statements for each task. You are required to implement your solution to each problem based on tasks' description.

2.1. Task 1 (65 marks)

You will be using clinical trial datasets in this work and combining the information with a list of pharmaceutical companies. You will be given the answers to the questions, for a basic implementation, for two historical datasets, so you can verify your basic solution to the problems. Your final submission will need to consist of results executed on the third, 2021, release of the data. All data will be available from Blackboard.

2.1.1. Datasets:

The data necessary for this assignment will be downloadable as .csv files. The .csv files have a header describing the file's contents. They are:

1. **Clinicaltrial_<year>.csv:**

Each row represents an individual clinical trial, identified by an Id, listing the sponsor (Sponsor), the status of the study at time of the file's download (Status), the start and completion dates (Start and Completion respectively), the type of study (Type), when the trial was first submitted (Submission), and the lists of conditions the trial concerns (Conditions) and the interventions explored (Interventions). Individual conditions and interventions are separated by commas.

(Source: ClinicalTrials.gov)

2. **pharma.csv:**

The file contains a small number of a publicly available list of pharmaceutical violations. For the purposes of this work, we are interested in the second column, Parent Company, which contains the name of the pharmaceutical company in question.

(Source: <https://violationtracker.goodjobsfirst.org/industry/pharmaceuticals>)

When creating tables for this work, you must name them as follows:

- clinicaltrial_2021 (and clinicaltrial_2019, clinicaltrial_2020 for the sample data)
- pharma

The uploaded datasets, must exist (and be named) in the following locations:

- /FileStore/tables/clinicaltrial_2021.csv (similarly for sample data)
- /FileStore/tables/pharma.csv

This is to ensure that we can run your notebooks when testing your code (marks are allocated for your code running).

You are to implement all steps 3 times: once in Spark SQL and twice in PySpark (For RDD and DataFrame).

For the visualisation of the results, you are free to use any tool that fulfils the requirements, which can be tools such as Python's matplotlib, Excel, Power Bi, Tableau, or any other free open-source tool you may find suitable. Using built-in visualizations directly is permitted, it will however not yield a high number of marks. Your report needs to state the software used to generate the visualization, otherwise a built-in visualization will be assumed.

2.1.2. Problem statement

You are a data analyst / data scientist whose client wishes to gain further insight into clinical trials. You are tasked with answering these questions, using visualisations where these would support your conclusions.

You should address the following questions. You should use the solutions for historical datasets (available on Blackboard) to test your implementation.

1. The number of studies in the dataset. You must ensure that you explicitly check distinct studies.
2. You should list all the types (as contained in the Type column) of studies in the dataset along with the frequencies of each type. These should be ordered from most frequent to least frequent.
3. The top 5 conditions (from Conditions) with their frequencies.
4. Find the 10 most common sponsors that are not pharmaceutical companies, along with the number of clinical trials they have sponsored. **Hint:** For a basic implementation, you can assume that the Parent Company column contains all possible pharmaceutical companies.
5. Plot number of completed studies each month in a given year – for the submission dataset, the year is 2021. You need to include your visualization as well as a table of all the values you have plotted for each month.

2.1.3. Extra features to be implemented for extra marks

If you answer correctly all 5 problems, you will get a “merit” mark for this task (maximum 69% percent of the task mark). You will get more marks If you implement several extra features like follows (but are not limited to):

- Maximum 3 further analyses of the data, motivated by the questions asked (new problem statements other than the 5 problems)
- Writing general and reusable code. For example, ensuring that switching to a different version of the data requires only the change of one variable.
- Using more advance methods to solve the problems like defining and using user defined functions.
- Successfully implementing Spark functions that you have not used in the workshop.
- Creation of additional visualizations presenting useful information based on your own exploration which is not covered by the problem statements.

2.2. Task 2 (35 marks)

2.2.1. Problem statement

For your second task, imagine you are working as a data scientist for a manufacturing company. They are using vibration sensors to monitor the machinery within their production line and want to use this data for predictive maintenance – the aim is to be able to classify whether there is a fault with the machine based on the readings from the vibration sensors. We have provided you with a dataset **FaultDataset.csv**. Each row contains twenty vibration sensor readings, and the final column identifies whether there was a fault with the machine at the time of the readings. In this column, 0 means there was no fault with the machine and 1 means a fault was identified.

When creating tables for this work, you must name them as follows:

➤ FaultDataset

The uploaded datasets, must exist (and be named) in the following locations:

➤ /FileStore/tables/FaultDataset.csv

This is to ensure that we can run your notebooks when testing your code (marks are allocated for your code running).

Your task as a data scientist is to do the following:

- Load the dataset into a Spark DataFrame. You may want to consider carrying out some initial exploratory analysis of the data, which you are welcome to do using DataFrames, Spark SQL, Databricks visualisations, another visualisation library etc.
- Use MLlib to train a Decision Tree classification model (DecisionTreeClassifier algorithm) on the provided data and evaluate its performance. You will need to carry out all pre-processing steps, such as splitting the data into training and test sets.
- Track your experiment with MLflow. You must include screenshots from the Databricks Experiment UI in your report to evidence that you have done this.

2.2.2. Extra features to be implemented for extra marks

For this task you will get more marks (more than 69%) If you implement several extra features like follows (but are not limited to):

- You can receive more marks for multiple runs as part of your experiment, for example, training models with different hyperparameters.
- You can get more marks for testing different classification algorithms and comparing the results. However, you should note that you are not being marked on your understanding of the different machine learning algorithms, as this is not an

assessed part of this module (this is covered in Machine Learning & Data Mining module).

3. Report structure

A 5000-word report that documents your solution should be included with your submission. **In this module, a background, literature review or citations are not required.** The format of the report should be as follows:

Task 1:

Compulsory Part

- 1) Description of any setup required to be able to complete this task
- 2) Data cleaning and preparation (including descriptions, justifications, and screenshots of all code)
- 3) Problem answers

Question 1

1. Assumptions made on the dataset before answering to this question
2. PySpark implementation outline in RDD (description of main ideas in words and screenshot of code)
3. PySpark implementation outline in DataFrame (description of main ideas in words and screenshot of code)
4. SQL implementation outline (description of main ideas in words and screenshot of code)
5. Discussion of result

Question 2

1. Assumptions made on the dataset before answering to this question
2. PySpark implementation outline in RDD (description of main ideas in words and screenshot of code)
3. PySpark implementation outline in DataFrame (description of main ideas in words and screenshot of code)
4. SQL implementation outline (description of main ideas in words and screenshot of code)
5. Discussion of result

Question 3

...

Optional Part

4) Further analysis 1

Use same structure as the main questions but only choose one implementation (RDD or DF or SQL)

5) Further analysis 2

Use same structure as the main questions but only choose one implementation (other than what you chose for the first further analysis)

6) Further analysis 3

Use same structure as the main questions but only choose one implementation (other than what you chose for the first and second further analysis)

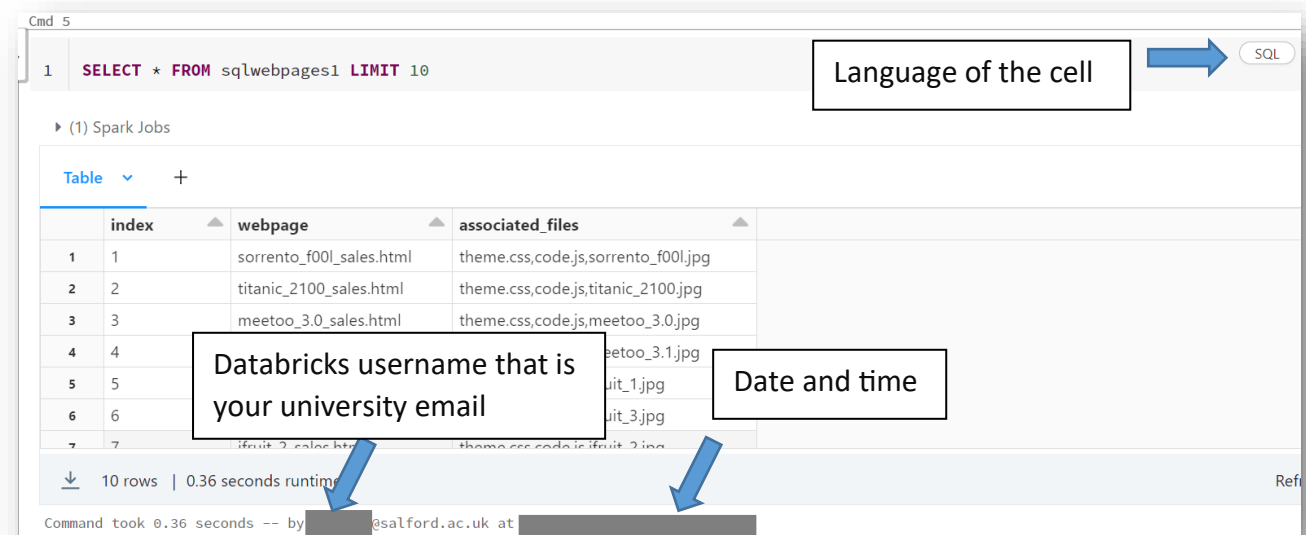
Task 2:

- 1) Description of any set up required to complete this task
- 2) Loading data into Spark DataFrame and any exploratory analysis or visualisation carried out prior to training
- 3) Data preparation and pre-processing carried out prior to training the model
- 4) Selection of hyperparameters and model training and evaluation and MLflow experiment tracking
- 5) Brief discussion of result

4. Important points to consider

Note 1:

All screen shots that you put in your report **MUST** have the metadata underneath of the image.



Language of the cell → SQL

index	webpage	associated_files
1	sorrento_f00l_sales.html	theme.css,code.js,sorrento_f00l.jpg
2	titanic_2100_sales.html	theme.css,code.js,titanic_2100.jpg
3	meetoo_3.0_sales.html	theme.css,code.js,meetoo_3.0.jpg
4		meetoo_3.1.jpg
5		uit_1.jpg
6		uit_3.jpg
7		

Databricks username that is your university email → salford.ac.uk

Date and time → 0.36 seconds running

Command took 0.36 seconds -- by [redacted]@salford.ac.uk at [redacted]

Without these metadata you will not get a mark for the implementation. If you need to refresh your skill for taking screenshots, [click here](#)

Note 2:

All your notebook **MUST** be rerunnable. You have learnt how to create a rerunnable notebook. This means the examiners will upload .zip datasets in their DBFS and when they press “Run All” button of your notebook then all the cells must be run without any error.

There are 2 main points that you should consider:

- 1) Don't change main datasets name
- 2) To make your notebook rerunnable you must write some scripts in a way that make sure your DBFS and SQL environment will be whipped out and ready to accommodate new databases, tables, views, dataframes, ... before each run.

- 3) For task 1, only your first notebook (<student name>_rdd.ipynb) must contains unzipping steps. Examiners have uploaded zipped files in their DBFS but your notebook should unzip data in their system and then run all your commands afterward.

Note 3:

For task 1, your solution must be implemented using both Spark SQL and PySpark.

Note that for the SQL implementation you **MUST** not use PySpark version of the SQL queries, and you must have a SQL notebook and such a SQL solution in PySpark will not receive any marks.

```

week 5 - lecture SQL
File Edit View Run Help Last edit was 15 days ago Give feedback
1
2 SELECT * FROM people WHERE pcode = 94020

```



```

week 5 - lecture Python
File Edit View Run Help Last edit was now Give feedback
Cmd 7
1 spark.sql("SELECT * FROM people WHERE pcode = 94020")

```



```

week 5 - lecture SQL
File Edit View Run Help Last edit was 2 minutes ago Give feedback
1 %python
2 spark.sql("SELECT * FROM people WHERE pcode = 94020")

```



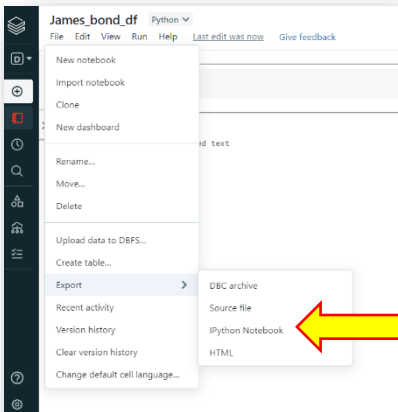
5. How to submit

You should submit your written report in the format of PDF document uploaded to Blackboard via the Turnitin submission area, with the name **<student-name>_report.pdf**
You should also submit your Databricks notebook in 3 formats.

1. IPython notebooks for RDD, DataFrame, and Machine Learning,
2. SQL notebook for the SQL part,
3. HTML format for all notebooks.

#	Content	Naming	Format	Example
1	report	Student name	pdf	James_bond_report.pdf
2	RDD implementation	Student name_rdd	iPython and html	James_bond_rdd.ipynb James_bond_rdd.html
3	DataFrame implementation	Student name_df	iPython and html	James_bond_df.ipynb James_bond_df.html
4	SQL implementation	Student name_sql	sql and html	James_bond_sql.sql James_bond_sql.html
5	Machine learning implementation	Student name_ml	iPython and html	James_bond_ml.ipynb James_bond_ml.html

Before saving your file in a html format run your notebook and make sure all outputs are shown and then save it.



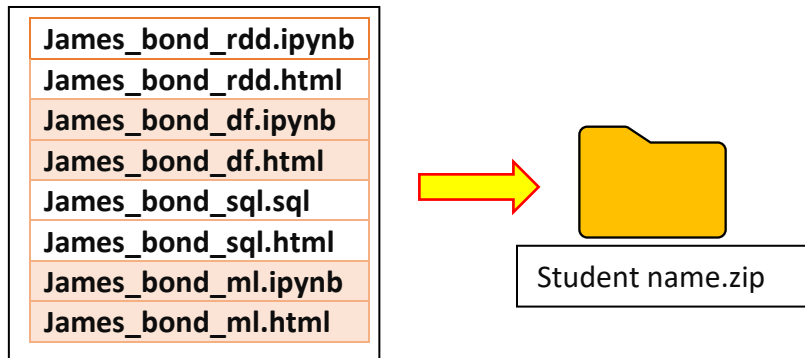
Use "Source file" to save as .sql

Use "IPython Notebook" to save as .ipynb

Use "HTML" to save as .html

You will see two submission areas on the blackboard, one for the pdf report and the second one just for codes.

You should put all codes (8 files) in a folder and then zip it and upload the zipped folder in the codes submission area.



6. Assessment Criteria

Level	Descriptor
Outstanding (90-100%)	<ul style="list-style-type: none"> • All the basic implementations of the task 1 and task 2 are correct • All the notebooks have been submitted in the proper formats as instructed in the assignment brief. • All the notebooks are rerunnable, and the code is clear, concise and fully commented • Use of extra features (e.g., user defined functions) to improve the code beyond the basic requirements of the brief • There are 3 further analyses that have been implemented correctly. • There are 3 extra features (other than further analysis) that have been implemented correctly. • For task 2, there is in-depth exploratory data analysis prior to training a machine learning model, which is clearly explained and justified in the accompanying report • For task 2, multiple machine learning experiments (at least 4) have been carried out and evaluated, and these are discussed in the report • Report structure is clear and is completely based on the assessment brief • Other than the report structure student has followed all the details that have been mentioned in the brief like naming the files, screenshots, ... • Throughout, the report demonstrates an excellent grasp of the underlying concepts and theory involved in the implementation, and all steps are explained in a rigorous and concise manner using correct terminology
Excellent (80-89%)	<ul style="list-style-type: none"> • All the basic implementations of the task 1 and task 2 are correct • All the notebooks have been submitted in the proper formats as instructed in the assignment brief. • All the notebooks are rerunnable, and the code is clear, concise and fully commented • Use of extra features (e.g., user defined functions) to improve the code beyond the basic requirements of the brief • There are 2 further analyses that have been implemented correctly. • There are 2 extra features (other than further analysis) that have been implemented correctly. • For task 2, there is in-depth exploratory data analysis prior to training a machine learning model, which is clearly explained and justified in the accompanying report • For task 2, multiple (at least 3) machine learning experiments have been carried out and evaluated, and these are discussed in the report • Report structure is clear and is completely based on the assessment brief • Other than the report structure student has followed all the details that have been mentioned in the brief like naming the files, screenshots, ... • Throughout, the report demonstrates an excellent grasp of the underlying concepts and theory involved in the implementation, and all steps are explained in a rigorous and concise manner using correct terminology
Very Good (70-79%)	<ul style="list-style-type: none"> • All the basic implementations of the task 1 and task 2 are correct • All the notebooks have been submitted in the proper formats as instructed in the assignment brief.

Assessment Information/Brief

Level	Descriptor
	<ul style="list-style-type: none"> • All the notebooks are rerunnable, and the code is clear, concise, and well commented • Use of extra features (e.g., user defined functions) to improve the code beyond the basic requirements of the brief • There are 1 further analysis that have been implemented correctly. • There are 1 extra features (other than further analysis) that have been implemented correctly. • For task 2, there is in-depth exploratory data analysis prior to training a machine learning model, which is clearly explained and justified in the accompanying report • For task 2, multiple (at least 2) machine learning experiments have been carried out and evaluated, and these are discussed in the report • Report structure is clear and is completely based on the assessment brief • Other than the report structure student has followed all the details that have been mentioned in the brief like naming the files, screenshots, ... • Throughout, the report demonstrates a very strong grasp of the underlying concepts and theory involved in the implementation, and all steps are explained in a rigorous and concise manner using correct terminology
Good (60-69%)	<ul style="list-style-type: none"> • All the basic implementations of the task 1 and task 2 are correct • All the notebooks have been submitted in the proper formats as instructed in the assignment brief. • All the notebooks are rerunnable, and the code is clear, concise, and well commented • For task 2, there is some exploratory data analysis prior to training a machine learning model, which is clearly explained and justified in the accompanying report • Report structure is clear and is completely based on the assessment brief • Other than the report structure student has followed all the details that have been mentioned in the brief like naming the files, screenshots, ... • Throughout, the report demonstrates a good grasp of the underlying concepts and theory involved in the implementation, and all steps are explained using correct terminology
Satisfactory (50-59%)	<ul style="list-style-type: none"> • Almost all the basic implementations of the task 1 and task 2 are correct • Almost all the notebooks have been submitted in the proper formats as instructed in the assignment brief. • Some of the notebooks are rerunnable • Report structure is reasonably clear and is primarily based on the assessment brief • Other than the report structure student has followed almost all the details that have been mentioned in the brief like naming the files, screenshots, ... • Throughout, the report demonstrates a reasonable grasp of the underlying concepts and theory involved in the implementation
Unsatisfactory (40-49%)	<ul style="list-style-type: none"> • Some correct implementations • All the notebooks are in proper formats as instructed in the assignment brief. • Notebooks are not rerunnable. • Most basic requirements of the brief have been met

Level	Descriptor
Inadequate (30-39%)	<ul style="list-style-type: none"> • Few correct implementations • Some basic requirements of the brief have been met
Poor (20-29%)	<ul style="list-style-type: none"> • Limited correct implementations • Some basic requirements of the brief have been met and datasets have not correctly been imported to the system
Very Poor (10-19%)	<ul style="list-style-type: none"> • No correct implementations for the tasks • Some basic requirements of the brief have been met and datasets have not correctly been imported to the system
Extremely Poor (0-9%)	<ul style="list-style-type: none"> • No attempt to correctly answer questions