XJTLU Entrepreneur College (Taicang) Cover Sheet

Module code and Title	**DTS303TC: Big Data Security and Analytics**
School Title	School of AI and Advanced Computing
Assignment Title	Individual Project (CW2)
Submission Deadline	**Wednesday, 26 October 2022, before 23:59 (China Time, GMT + 8)**
Final Word Count	N/A
If you agree to let the university use your work anonymously for teaching and learning purposes, please type **"yes"** here.	

I certify that I have read and understood the University's Policy for dealing with Plagiarism, Collusion and the Fabrication of Data (available on Learning Mall Online). With reference to this policy I certify that:

- My work does not contain any instances of plagiarism and/or collusion.
- My work does not contain any fabricated data.

By uploading my assignment onto Learning Mall Online, I formally declare that all of the above information is true to the best of my knowledge and belief.

	Scoring – For Tutor Use
Student ID:	

Stage of Marking	**Marker Code**	**Learning Outcomes Achieved (F/P/M/D) (Please modify as appropriate)**					**Final Score**
				C	**D**		
1st Marker – red pen							
Moderation – green pen	**IM Initials**	The original mark has been accepted by the moderator (please circle as appropriate):					Y / N
		Data entry and score calculation have been checked by another tutor (please circle):					Y
2nd Marker if needed – green pen							

For Academic Office Use			**Possible Academic Infringement (please tick as appropriate)**	
Date Received	**Days late**	**Late Penalty**	☐ **Category A**	Total Academic Infringement Penalty (A, B, C, D, E - Please modify where necessary) ______________________
			☐ **Category B**	
			☐ **Category C**	
			☐ **Category D**	
			☐ **Category E**	

Module	EXAMINER	DEPARTMENT	Email
DTS303TC	Md Maruf Hasan	School of AI and Advanced Computing	MdMaruf.Hasan@xjtlu.edu.cn

1st SEMESTER 2022/23 Individual Project (CW2)

Undergraduate – Year 4

DTS303TC: Big Data Security and Analytics

Submission Deadline: Wednesday, 26 October 2022, before 23:59 (GMT+8)

INSTRUCTIONS

1. The weighting of this INDIVIDUAL project is 60% of the final mark.
2. The detailed guideline including the marking scheme is provided below.
3. Your submission should only be in English and your submission should be according to the instructions provided for each question/section below.
4. Please use Microsoft Word to typeset your document professionally. The report must be submitted in PDF format (a single file) via Learning Mall Online to the correct dropbox. The report should answer the questions asked with evidence of the task performed.
5. Only electronic submissions are accepted, and no hard copy submission is required. You should include code snippets, detailed steps and screenshots appropriately to show the evidence of your work.
6. All related code files should be submitted separately as a single zip file. Please include a README file with instructions about how to execute the code if appropriate.
7. After final submission, all students must download their file and check that it is uploaded correctly and viewable. Documents may become corrupted during the uploading process (e.g., due to slow internet connections). However, students themselves are responsible for submitting a functional and correct file for their assessments.
8. Late submission penalty applies according to the university policy

Student ID	**Student Name**
DEPARTMENT	**Email**

DTS303TC: Individual Project, CW2 (60%)

The Project Overview:

When we use publicly available datasets to practice basic data exploration, analytics and visualization in a lab environment, we do not have to worry much about data governance, privacy and security. The default installation setup of big data clusters like those powered by Hadoop and Spark families do not address the privacy and security concerns adequately. For example, the default setup for Hadoop doesn't include encryption for both data-at-rest and data-in-transit. Hadoop includes an access control and authorization mechanism of its own. However, in an enterprise big data deployment, it is preferable that the Hadoop cluster integrates the enterprise Active Directory or Kerberos-based authentication and authorization to enforce IT governance including audit. In the second half of this module, students will learn about basic data security principles, techniques and tools to ensure privacy and compliance from an IT governance perspective which is essential to any big data deployment.

Be it on-premise, cloud or a hybrid architecture, in a real-life big data deployment, we need to comply with data governance policy and address privacy and security concerns thoroughly. In this project, students will start with a small and manageable dataset to practice some simple data exploration and visualization tasks (PART A), followed by a realistic machine learning task (PART B). As the student familiarize themselves with this publicly available dataset, they are asked to consider a realistic big data deployment with additional data items that are likely to enter into a data warehouse (data lake) to support real-time business functions in data-driven decision-making and service development that comply with privacy and security objectives of an organization (PART C).

While PART A and B can be tried easily on a local installation of Hadoop/Spark, students are expected to utilize the Sugon Platform of the university or have a single-node (pseudo-distributed) installation of Hadoop on a single computer to try out PART C during the lab sessions to gain first-hand experience on Big Data Security.

Upon completion of PART A, B and C, students will prepare a comprehensive report (a PDF file) based on the guideline given below to reflect the following module learning outcomes. Students should also submit their code and script files together with a README file (a ZIP file) with instructions to execute/verify their work.

Learning Outcomes:

This coursework will be assessed for the following two learning outcomes:

C. Show *proficiency* with at least one data analytics software package.

D. Demonstrate *awareness* of issues related to computer and data security

The Dataset:

Explore the dataset of the **New Car Sales in Norway**. The dataset is publicly available and downloadable from the following link:

https://www.kaggle.com/dmi3kno/newcarsalesNorway.

The dataset contains monthly car sales for 2007–2017 by make and most popular models. The Dataset includes three csv files as follows:

1. Norway_new_car_sales_by_make.csv:
Monthly sales of new passenger cars by make (manufacturer brand)

1. Year – year of sales
2. Month – month of sales
3. Make – car make (e.g. Volkswagen, Toyota, Tesla)
4. Quantity – number of units sold
5. Pct – percent share in monthly total

2. Norway_new_car_sales_by_model.csv:
Monthly summary of top-20 most popular models (by make and model)

1. Year – year of sales
2. Month – month of sales
3. Make – car make (e.g. Volkswagen, Toyota, Tesla)
4. Model – car model (e.g. BMW-i3, Volkswagen Golf, Tesla S75)
5. Quantity – number of units sold
6. Pct – percent share in monthly total

3. Norway_new_car_sales_by_month.csv:
Summary statistics for car sales in Norway by month

1. Year – year of sales
2. Month – month of sales
3. Quantity – total number of units sold
4. Quantity_YoY – change YoY in units
5. Import – total number of units imported (used cars)
6. Import_YoY – change YoY in units
7. Used – total number of units owner changes inside the country (data available from 2012)
8. Used_YoY – change YoY in units
9. Avg_CO2 – average CO2 emission of all cars sold in a given month (in g/km)
10. Bensin_CO2 – average CO2 emission of bensin-fueled cars sold in a given month (in g/km)
11. Diesel_CO2 – average CO2 emission of diesel-fueled cars sold in a given month (in g/km)
12. Quantity_Diesel – number of diesel-fueled cars sold in the country in a given month
13. Diesel_Share – share of diesel cars in total sales (Quantity_Diesel / Quantity)
14. Diesel_Share_LY – share of diesel cars in total sales a year ago
15. Quantity_Hybrid – number of new hybrid cars sold in the country (both PHEV and BV)
16. Quantity_Electric – number of new electric cars sold in the country (zero emission vehicles)
17. Import_Electric – number of used electric cars imported to the country (zero emission vehicles)

The Norway new car sales dataset can be used for analyzing and predicting future car sales. Explore the dataset to solve the following problems for analysis, prediction and visualization using **Sklearn** and **PySpark**.

Students may use other big data software. However, detailed steps including codes and outputs are to be submitted together with the report so that the answers can be reproduced or verified easily.

PART A – DATA EXPLORATION – 30% (Outcome C – Data Analytics Software)

[Choose any 10 tasks; 10 x 3 marks each; 30 Marks]

1. Print year-wise total car sales and visualize the output (Hint: use bar chart for Year vs. total car sales).

2. Print monthly total car sales and visualize for a specific year.

3. Print monthly total car sales from 2007 to 2017 and visualize them to represent the month numbers (1 for Jan 2 for Feb) and total car sales value. (Hint: Use bar chart). Also, find the month for number of highest and lowest car sales.

4. Calculate the total amount of the sales for each manufacturer from 2007 to 2017. Find the top 10 manufacturers based on the total sale and visualize the output. (Hint: Sort make-wise total car sales and visualize them using bar chart).

5. Rank top 10 car brands. Visualize the year-wise result using line graph

6. Draw pie chart for the sales of all the models of "Toyota" in year 2012.

7. Find which model of each manufacturer has the highest sales in year 2015.

8. Find which model of each manufacturer has the highest sales during 2007 to 2017.

9. Find which model of each manufacturer has the lowest sales during 2007 to 2017

10. Compare car models with percentage shares.

11. Plot the sales of new cars and sales of the diesel cars to see the comparison. Similarly, plot the sales of new car and electric cars. (Hint: Use line chart).

12. Calculate year-wise share of diesel car sales in total sales.

13. Compare year-wise average consumption of CO2 emission of all cars sold with year-wise average consumption of CO2 emission in benzene-fueled cars sold and diesel-fueled cars sold.

14. Calculate and visualize year-wise new and used (import) car sales to compare the statistics.

15. Calculate and visualize year-wise sales of all used (import) car and sales of electric-used cars (import_electric) to make a comparison.

What to submit:

For each question, explain your approach and include the code or detail steps followed by the output data and visualization.

Grading Scheme:

Code/step and analysis: 2 mark

Correct/appropriate output and visualization: 1 mark

PART B – Machine Learning – 30% (Outcome C – Data Analytics Software)

[Choose any **one** of the following prediction tasks. 30 Marks]

1. Predict (forecast) and visualize the car sales for all the months of 2020 using the month-wise car sales quantity from the Jan 2007 to Jan 2017.

2. Predict (forecast) and visualize the month-wise rise of green vehicles in 2018. (Hint: Green Vehicle = Import_Electric + Quantity_ Hybrid + Quantity_Electric).

3. Predict (forecast) and visualize the diesel market share in 2019. State whether it represents growth or reduction in the sales.

What to submit:

Individual students will prepare the answer to the question they choose. Explain your approach with justifications. Include the code snippets or detail steps with intermediate data/results that clearly shows your understanding of the machine learning approach you used so that training, testing, validation and parameter tuning etc. are clearly understandable in your report.

Visualize and comment on your intermediate steps/results as much as possible. Include a convincing discussion to justify your prediction result and its strengths and weaknesses.

It is recommended that students will use a robust forecasting model and identify and address the seasonal components or other intricate aspects in the dataset.

Grading Scheme:

A. Selection of appropriate machine learning techniques with justifications: 10 Marks
   - Thorough analysis and justification of the approach and algorithm used with consideration such as seasonal components: 8-10
   - Appropriate use of built-in algorithms that produce accurate/acceptable results: 4-7
   - Partially correct answers with some justification: 0-3

B. Appropriate training, testing and validation: 10 marks
   - Detail description of the training, testing and validation including justification of parameter tuning: 8-10
   - Acceptable descriptions of training, testing and validation: 4-7
   - Partially correct answers: 0-3

C. Conclusion (discussions and justification of the overall prediction): 10 marks
  - Valid analysis of the strengths and weaknesses of the prediction results according to the machine learning approach used with comparisons and references made to other alternative approaches: 7-10
  - Acceptable analysis of the strengths and weaknesses of the prediction results according to the machine learning approach used: 4-6
  - Poor justifications and lack of proper understanding of the machine learning approach used: 0-3

PART C – DATA SECURITY – 40% (Outcome D – Awareness of Data Security)

An essay on Compliance, Privacy and Security Considerations for Big Data

The given publicly available dataset above is prepared after removing sensitive data and information. We may use them freely for data analytics and machine learning demonstration without having to worry about privacy, security and compliance issues. However, in real-life Big Data applications, huge amount of inter-related data from heterogeneous sources are continuously accumulated and stored in a distributed cluster where security, privacy and compliance requirements are paramount. The access and use of big data must comply with the privacy and security-related policy and law.

You may consider the Ministry of Transport or the Environment Protection Agency of a country may regularly receive data from all car manufacturers, dealers and customers that may include Personally Identifiable Information (PII) or other sensitive information. The It is not desirable that every employee or user is authorized to access every piece of data. Authentication, authorization and encryptions etc. must be enforced based on the organization's policy as well as national and international law. To make real-time Big Data -driven decision-making, government agencies or business entities must abide by the legal and ethical guidelines in storage, transmission and processing of such data (e.g., you may not be allowed to store certain data on a server outside your national boundary!).

In this part of your project, you will introduce additional data items (in the given CSV files or add additional CSV files) with some mock data that includes PII (Personally Identifiable Information) or other sensitive data about car sales and ownerships (e.g., Customer, Dealer and Repairing Workshops, Payment and Financial Information etc.) and upload the combined dataset on a Big Data Ecosystem with proper consideration of Compliance, Privacy, and Security issues as if they are ready to deploy.

If the Sugon Platform is available for use, students are advised to utilize the Sugon Cluster. Alternatively, they may use a Single Node (Pseudo-distributed) Hadoop installation on the lab or personal computer. It is also possible to try things out on a commercial cluster platform as free/trial cloud services and platforms are offered by Cloudera, MongoDB, Databricks, Amazon, etc.

Please note that in PART C, you are only expected to demonstrate the data security measures from a proof-of-concept perspective rather than a complete deployment. Hence you will write a short essay describing security and privacy ramification together with the tools and techniques you will adopt to ensure big data security in your chosen platform or scenario.

Note: Due to the scheduled campus relocation, the XJTLU Sugon Platform may be unavailable temporarily.

What to submit:

You will submit **an individual essay** (recommended length: 2,000 words or less) that include the following.

1. Descriptions of all the data items and relevant compliance, privacy and security consequences in a realistic context. (10 Marks)
2. Description of the overall architecture of the Big Data Ecosystem. Detail description of the Big Data Ecosystem with necessary layers and components using a block diagram is desirable. Students must also highlight the relevant compliance, privacy and security issues and how they are planned to be mitigated in their deployment. (20 Marks)
3. Conclusion and discussion referring to similar real Big Data deployment case studies and references as discussed in the lecture. (10 Marks)

The essay should reflect a student's individual understanding of the big data ecosystem in terms of technology as well as security, privacy and compliance considerations.

All data items (including the newly introduced PII data) should be stored on the cluster platform with proper security and privacy considerations similar to a real deployment. The diagram must include basic Hadoop and Spark components as well as those you used (or recommend using) to ensure security and privacy objectives. In particular, the reports should include most of the following considerations and evidence on your cluster setup and deployment using Hadoop and Spark suite:

- Creating a cluster with SSL support using Digital Certificates
- Key Management strategy and enabling SSL for HDFS, SPARK and YARN
- Authentication setup using Kerberos or Active Directory integration
- Encryption setup for HDFS or Spark Encryption Zone for both data at rest and data in transit (e.g., TLS)
- Role-based Authorization using Sentry and HDFS Access Control List

See the **Grading Rubric** in the table below:

Category	Exemplary (16-20 Points)	Accomplished (12-15 Points)	Capable (8-11 Points)	Need Improvement (0-7 Points)	ILOs	Points	Weight	Marks Received = Points*weight
Description of the Data items in the right context with proper Security and	*Descriptions of all the sensitive data items and relevant compliance, privacy and security consequences in appropriate contexts with proper justifications. References are also made to the relevant regulations and compliance framework.*	*Descriptions of all the sensitive data items and relevant compliance, privacy and security consequences in appropriate contexts with proper justifications.*	*Descriptions of most of the sensitive data items and relevant compliance, privacy and security consequences in appropriate contexts.*	*Descriptions of some the sensitive data items and relevant compliance, privacy and security consequences in the right context*	D	20	0.5	
Description of the overall architecture and deployment with relevant tools/technologies	*Description of the overall architecture of the Big Data Ecosystem is presented clearly with a comprehensive block diagram with appropriate selection of components and tools. Proper justifications and comparisons are added. References are also made to the relevant technologies and components using proper reference and use-cases. Advanced topics such as corporate and national policy, data aggregation and de-anonymization issues as well as security and privacy concerns introduced by machine learning itself are also highlighted*	*Description of the overall architecture of the Big Data Ecosystem is presented clearly with a comprehensive block diagram with appropriate selection of components and tools. Proper justifications and comparisons are added. Considerations for Encryption, Access Control, Auditing, Vulnerability Analysis and Threat Modeling are identified, and appropriate measures are implemented or recommended*	*Description of the overall architecture of the Big Data Ecosystem is presented clearly. A comprehensive block diagram with necessary layers and components is also presented with appropriate selection of components and technologies.*	*Description of the overall architecture of the Big Data Ecosystem is somewhat present. However, a clear and comprehensive block diagram with necessary layers and components is not presented.*	D	20	1	

Conclusions and Discussions with proper references and use cases	*All security and privacy implications are clearly identified in the context of a realistic deployment of Big Data application using a scalable, reliable and secure platform with necessary tools and components. Proper justifications and comparisons are added. Alternative choices are considered as well as proper references are made to relevant use cases and references including national and international standards and best practices.*	*All security and privacy implications are clearly identified in the context of a realistic deployment of Big Data application using a scalable, reliable and secure platform with necessary tools and components. Proper justifications and comparisons are added that shows significant understanding of the technologies and tools.*	*Most security and privacy implications are clearly identified in the context of a realistic deployment of Big Data application using a scalable, reliable and secure platform with necessary tools and components.*	*Some security and privacy implications are clearly identified in the context of a realistic deployment of Big Data application using a scalable, reliable and secure platform.*	D	20	0.5	
TOTAL MARKS = 40				**Marks Obtained**				

[END]