AIP PROJECT

Program Name: Big Data Analytics

Project Code: CPL-5559-DSMM

AIP Project Case Study Orientation

Vocational Learning Outcomes (VLOs) Covered in this WIL Project Case Study

1. Collect, house, extract, manipulate, maintain, and mine data sets that respond to organizational, financial, or market needs.
2. Recommend different systems and network architectures, artificial intelligence, and data storage technologies to support data analytics and Big Data.
3. Design and apply data models that meet the needs of a specific operational process or business model.
4. Develop software applications, algorithms, and artificial intelligence models to manipulate, correlate and reduce data sets and produce project documentation and reports.
5. Collaborate effectively with diverse teams to design and present data visualizations that communicate Big Data concepts and information to stakeholders and business professionals.
6. Apply business analytics, business intelligence tools and research to support evidence-based decision-making that helps an organization meet their stated objectives.
7. Identify and assess data analytics and Big Data business strategies and workflows to respond to new opportunities or provide project solutions.
8. Implement data security solutions in compliance with corporate security policies, ethical standards, and industry regulations.
9. Deliver data-oriented projects using data science, business analysis, and project management principles, tools, and techniques.
10. Develop artificial intelligence solutions to support administration, decision-making, planning, risk management, logistics, manufacturing, smart devices, and robotics.

Essential Employability Skills (EESs) Covered in this AIP Project Case Study

- Communication
  - It helps to communicate clearly, correctly, and concisely in different forms. These include oral, written, and visual.
- Numeracy
  - This skill set helps to solve mathematical operations effectively with accurate precision.
- Critical thinking & problem solving

  - It is a systemic approach to attempting to resolve problems by analyzing the pros and cons of a decision.
- Information management
  - It helps to locate, select, organize, and document information with the use of technology by analyzing aspects and gathering information from a variety of sources.
- Interpersonal Skills
  - This skill set is important as it helps to respect others' opinions or input. It helps to build teams and maintain relationships to achieve overall team or organizational goals.
- Personal Skills
  - These soft skills are important in developing employability talents, such as dependability, adaptability, and problem-solving skills.

AIP PROJECT

Program Name: Big Data Analytics

Project Code: **CPL-5559-DSMM**

Week 1

This Week's Detailed Case study Information

Congratulations you have been hired as a Data Engineer for Deepsight Analytics located in Mississauga ON.

Deepsight Analytics is a boutique consulting company focusing on data analytics and predictive modelling in some of the high-growth areas such as finance, retail, social media, and online advertising. It has a strong team of Data Scientists, Data Engineers, and Business consultants with in-depth knowledge and experience in solving data-related problems such as social media analytics, customer insights, customer analytics, forecasting, inventory and promotion planning, scheduling, pricing and revenue management and fraud detection.

The company has been storing the day-to-day transactional data of credit cards for a few years now. The transactional data are collected from multiple sources such as payment channels, retail, and eCommerce. All these data are stored in the distributed environment with proper replication factors. The Hadoop ecosystem has been used to store and manage those data.

The CEO of Data Analytics, Dr. Brian Campbell, wrote a memo to all employees and teams that the 21st economy is driven by data and, therefore, data analysis, data management, data synthesis and data visualization are some of the top investment strategies of many multinational corporations, non-profit organizations and governments. Without Information Technologies, there is no business success. However, our understanding of data has changed; more than 15 years ago, data had a predictable format, storage and analytics. Since the 2000s, our world has become more globalized and interconnected with the development of social media and new digital apps. These innovations force businesses to change their traditional methods of data analysis since data are now complex and multidimensional. Data analysis is at the heart of this process because the extraction of meaningful information is needed to determine what counts as data, and how the new data can be utilized for various business purposes. In the early 2000s, the key challenge was data storage and processing. In 2005, Hadoop ecosystems were introduced into the market. Now, we have no choice but to become the leading data expert in our changing and innovative 21st-century economy.

This memo is presented at your first team meeting during your onboarding week. Your data manager, Hannah Clark, introduced the memo to all team members. In her introduction, she highlighted that our department has to build a fraudulent transactions detection system because there have been many reported cases of fraudulent transactions in major credit card systems. The team is supposed to use the components of the Hadoop ecosystem HDFS, Kafka, Hive, Sqoop, Spark and Scala to gather the data from multiple partitions, transform the data, perform Exploratory Data analysis (EDA) and build the model to classify the fraudulent transactions. Hannah is adamant that the team has to rely on the F1 score as a metric to evaluate the algorithm. She said that the team will need to observe the performance of the different models and gather the reason for choosing a particular algorithm. Everyone seems to be very excited to work on this development. At the end of the meeting, Hannah introduced you to the team:

- Francis Underwood (Lead Data Scientist)
- Logan Arnold (Data Scientist)
- Hanna Clark (Data manager)
- Claire Hale (Project Manager)
- Dhaval Patel (Data Engineer)
- Amandeep Singh (Data Engineer)
- Andrew NG (Data Analyst)
- James Howlett (Data Analyst)

You will be working closely as a Data Engineer with a few of the team members, such as Amandeep Singh, Andrew NG and James Howlett.

Francis is a Data Scientist and has been working in Finance domain for 10 years now. He has provided expert knowledge in multiple projects related to Finance. Before working for Deepsight Analytics, he used to work with Banks in the credit department. Claire is the project manager and has been working on different types of projects. She used to manage software development projects and now she is managing Data Analytics projects for Deepsight Analytics. Dhaval is Data Engineer his expertise is in python, excel and, SQL. He takes care of Data pipelines and prepares data for predictive and prescriptive modelling. You feel that you will learn and grow as a member of this multi-professional team. You can't be more excited about your onboarding week and the memo from Dr. Campbell. Clearly, you now can apply your previous knowledge in a real-life context of innovation in the domain of data engineering.

Week 1 Onboarding Expectations and Participation

Your task this week is to participate in training and orientation for DeepSight Analytics. You will participate in a variety of exercises that are designed to get to know you better and understand your role within the team. You will participate in Team-building exercises that

prepare you for success within the Project. As with any position, you will have an excellent opportunity to build on your skills as a leader so long as you put forth your best effort. Use this week to develop a communication plan with your team and be ready to dive into the deliverables starting next week.

Note: You can make any assumptions that are deemed necessary for each case on a week-by-week basis. You will not be provided direct answers or 100% of the information necessary to complete each deliverable. Instead, focus on delivering the highest quality outcome that you can, as a group, to highlight your talent. You would be presenting these deliverables to Claire and would want to ensure that the work is of the highest quality.

This section will be available to you for the entirety of the project. However, each subsequent week's case study information may only be available for that week. Be sure to download and save this week's information for future use.

Week 2

Applicable VLOs or EESs for This Week's Case Study

1. Collect, house, extract, manipulate, maintain and mine data sets that respond to organizational, financial, or market needs.

2. Recommend different systems and network architectures, artificial intelligence, and data storage technologies to support data analytics and Big Data.

7. Identify and assess data analytics and Big Data business strategies and workflows to respond to new opportunities or provide project solutions.

This Week's Detailed Case Study Information

Hannah scheduled a team meeting for Francis, Claire, Dhaval and you. There will be a discussion about the tools and stacks that will be used in this project. The use of the tools and stack depends on the nature of the problem, the skillset of the team and the budget of the project. Before attending the meeting, a detailed market analysis regarding the issues is always helpful. In this project, there are lots of tools that can be leveraged to provide the solution. Perform thorough market analysis and document all of the stacks that are ideal for this use case. After the discussion with all of your team members, the tools and technology to be used for this project will be decided. As a Data Engineer, you should be up to date with emerging technologies their alternatives and their limitations. Hence, you should prepare for the meeting to ensure that you present yourself professionally and with confidence.

Deliverables for This Week's Case Study

1. Research and understand the basic mechanism of how Hadoop works.
2. Compare Legacy Hadoop Systems and modern solutions for Big Data. Consider the following topics for the basis of comparison:
   - Budget
   - Use Cases
   - Scalability
   - Robustness
   - Speed
   - The working mechanism of the underlying architecture
3. Point out the challenges which were faced by the legacy Hadoop system and solved in modern solutions.

AIP PROJECT

Program Name: Big Data Analytics

Project Code: **CPL-5559-DSMM**

Week 3

Applicable VLOs or EESs for This Week's Case Study

2. Recommend different systems and network architectures, artificial intelligence, and data storage technologies to support data analytics and Big Data.
3. Design and apply data models that meet the needs of a specific operational process or business model.
4. Develop software applications, algorithms, and artificial intelligence models to manipulate, correlate and reduce data sets and produce project documentation and reports.

This Week's Detailed Case Study Information

This week Dhaval told you that OLAP (online analytical Processing) is a computing method that enables users to extract data easily and selectively to analyze it and draw insights. Usually, credit card transactional data are written once and read multiple times. These data stored over time can be used for analytics purposes.

From your conversations with Dhaval, you learned that Deepsight Analytics has stored the data in HDFS (Hadoop Distributed File System) in CSV format. The data are collected from multiple variant sources and written in HDFS. These data are mostly meant for reading accesses. Not all analytical processes have the same data requirements, data requirement totally depends on the scope of the project and business needs.

Only after defining the scope of the project, the team will have a clear picture of data requirements. Before performing any transformation and analytical processes, the scope of the project should be documented.

In the ideal case, each and every task will be completed in time and the team will be able to meet the specified deadline and reach milestones but if the team runs into any sort of problem which might delay the task completion it is better to inform Claire early so that she can reassess the situation and plan accordingly.

You should attend a very important meeting on Thursday with Dhaval (DE), Claire (PM) Andrew NG (Data Analyst), stakeholders and other concerned parties to finalize the scope of the project. Claire will be presenting some of the tentative deadlines, milestones and deliverables but they will be finalized only when the whole team agrees on the timeline for each task and deliverable.

Claire would like your team to use agile methodologies, and each sprint will have either weekly or biweekly deliverables to be submitted to Hannah. Constructive feedback from every week will be documented so that the next sprint can be planned accordingly. All the team members will be there at weekly stand-up meetings to discuss the task and problems faced by the team and find a solution to them. These meetings will be short and effective so that the team can focus more on their daily tasks.

Deliverables for This Week's Case Study

1. Make sure all the required applications are installed and configured to implement this use case
2. Install Hadoop, Hive
3. Identify the different file formats to be used for this use case
4. Create a sample data file to send to HDFS
5. Write a python script to convert the parquet file to CSV and test it.
6. List commands used for reading and writing data to HDFS

Week 4

Applicable VLOs or EESs for This Week's Case Study

4. Develop software applications, algorithms, and artificial intelligence models to manipulate, correlate and reduce data sets and produce project documentation and reports.

This Week's Detailed Case Study Information

Claire wants you to identify the triple constraints for the project (Time, scope and cost). Before doing that, you need to have a clear picture of the task involved in the project. First, identify the major milestones of the project and allocate time to each milestone. Since it's a machine-learning problem, you need to take care of each step in the ML workflow and allocate time for it. As mentioned in earlier weeks, an agile team including Francis, Claire, Dhaval, Hannah, Andrew and you would be the human resource for the project. The project will last for about four months. Since you have the estimates of the resource and time you can prepare the budget for the project. Once you are done with the timeline estimate the budget for the project.

Deliverables for This Week's Case Study

1. Action Plan - progress-to-date
    - Estimate the total budget of the project that DeepAnalytics must pay
    - Document outlining time, scope and cost
2. Determine the short-term goals and transform the strategies you make in the last week into a specific performance target.

AIP PROJECT

Program Name: Big Data Analytics

Project Code: **CPL-5559-DSMM**

Week 5

Applicable VLOs or EESs for This Week's Case Study

4. Develop software applications, algorithms, and artificial intelligence models to manipulate, correlate and reduce data sets and produce project documentation and reports.

This Week's Detailed Case Study Information

Credit Card data have billing addresses, names, and other personal information of the users. All this information might not be needed for analytics processes. Before providing the data to your analytics team, it is a good practice to go through metadata and filter out the columns which are not needed to achieve the project objective. There are several ways to preserve the confidentiality of the data, data removal and encryption, data coarsening and data masking, and Principal component Analysis (PCA) are a few of them. Claire asked to find an ideal method to mask the data set containing the features mentioned earlier. The ideal method will not be computationally heavy and get the work done easily. Before using any python library, try to understand the use case of those methods. There are multiple data masking techniques, but you should choose the one which suits your use case. Be ready to explain the reason behind using a particular technique. You can always refer to the documentation of those techniques to learn more about them and justify your use case.

Before shipping data to a different location or performing any kind of transformation, the metadata(data about data) should be clear. Without knowing the metadata or structure of the data, no operation should be performed. To create a table in the hive you need to know the structure of the data. The data may be zipped or raw. There are different possible file formats and before diving into ETL processes make sure you are clear about the file formats and structure.

Deliverables for This Week's Case Study

- Explore the format of your data and identify the required field
- Analyze the data type required for each field
- Research the use case scenario of the external and internal table in hive.
- Point out the difference between the internal and external tables in the hive.
- Create an external table in the hive and load the dataset from HDFS to Hive
- Point out the difference between the internal and external tables in the hive

Week 6-Midterm Week

Mid-Term Panel Evaluation Preparation

Hannah asked your team to present the work in progress in front of Dr. Campbell, external investors and prospective clients. This presentation is crucial because the panel will evaluate your work and will provide your team with constructive feedback. You will need to demonstrate your professional work, soft and hard skills as well as leadership competencies. Your presentation should be innovative, creative and multimodal highlighting the key aspects of your project. Dhaval proposed using Canva software for the development and design of your panel presentation, but you are not sure if it is the most effective method. In fact, you have never utilized this software before. You told Dhaval that a Microsoft PowerPoint may be the best choice because it is simple and user-friendly; it is also safe for you because you know how to use it. Dhaval is becoming too emotional about Canva and insists on using it. The clash of ideas makes you feel scared and anxious about the presentation; you try to remain calm and improve your emotional intelligence by analyzing the pros and cons of Dhaval's proposal.

However, Dhaval thought that your doubts undermine your team cohesion and decided to discuss this clash with Hannah. Clearly, you should improve your problem-solving skills and leadership competencies. Hannah is concerned with the lack of team collaboration and the lack of soft skills in your team. She expressed her concerns via email suggesting that you should learn more about Canva in order to understand Dhaval's proposal and to create a fruitful dialogue with him. Hannah said that you should be able to step outside of your comfort zone and, in doing so, you should solve your challenge in the workplace. Hannah concluded that she trusted your team's decision and solution. You decide to follow Hannah's advice because it is in your and your team's best interests. That is why you should learn more about Canva and collaborate with Dhaval; your team should work together and create the best presentation using the best media and technology available at your disposal.

Presentation and Oral Delivery

CONTENT

- Overview of work in the Deep Sight Analytics
- Highlight three key areas you find of interest: Two areas related to weekly work completed and one area to highlight your learning from professional development sessions
- Apply reflecting skills
- Present the importance/benefit of your work to DeepSight Analytics.

Program Name: Big Data Analytics

Project Code: CPL-5559-DSMM

Week 7

Applicable VLOs or EESs for This Week's Case Study

8. Implement data security solutions in compliance with corporate security policies, ethical standards, and industry regulations.

7. Identify and assess data analytics and Big Data business strategies and workflows to respond to new opportunities or provide project solutions.

This Week's Detailed Case Study Information

Assessing the risk of a project is part of project planning. Deepsight analytics has a strict policy of providing the data on a "Need to Know" basis to reduce the risk of data privacy violation. Claire has requested you analyze, the datasets carefully and remove the unnecessary fields before providing them to the analytics team. Be proactive and research how PCA can help in data masking, identify the risks involved such as low-quality data and improper analytics in the project and the severity of the risk.

Introduction to this week's task

Week 7- Risk Analysis

1. Risk identification and mitigation strategies This week, your team will identify the potential risks associated with your solutions or idea. Using the Risk matrix conduct risk analysis and find the possible risks faced by the company because of performing Credit card transaction analysis and the weakness pointed out earlier in weeks 4 and 5 using SWOT Analysis what are the required strategies to mitigate them?
2. Software/hardware methods, techniques and tools needed. Moreover, based on the above research provide Software/hardware methods, techniques and tools that might be required for risk mitigation.

Deliverables for This Week's Case Study

1. Create a tabular detailed work highlighting risk analysis using Risk Matrix indicating risks, and vulnerabilities as per your solution or hardware used.

S.no	Risks	Type/Domain	Level (Low/Medium/ High)

Technical Deliverables

1. Identify any security issues.
2. Propose security tools to prevent any security-related issues.
3. Consider a different number of features while performing PCA and compare the results.
4. Perform research in Role Based Access control (RBAC)

AIP PROJECT

Program Name: Big Data Analytics

Project Code: **CPL-5559-DSMM**

Week 8

Applicable VLOs or EESs for This Week's Case Study

3. Design and apply data models that meet the needs of a specific operational process or business model.

4. Develop software applications, algorithms, and artificial intelligence models to manipulate, correlate and reduce data sets and produce project documentation and reports.

10. Develop artificial intelligence solutions to support administration, decision-making, planning, risk management, logistics, manufacturing, smart devices, and robotics.

This Week's Detailed Case Study Information

Deepsight analytics has outsourced the data labelling task, so now it has labelled data identifying the fraudulent transaction. Now the dataset for training and testing purposes is available. The next step is to initiate ML workflow.

Based on the datasets, Francis asked you to build a model that will identify the fraudulent transaction from the genuine ones. Francis is skeptical about your skill sets for building models for AI/ML workflows. Prove him wrong by performing all the steps of the machine learning workflow and building the model that will identify the fraudulent transaction. Use a confusion matrix for evaluating the performance of the model and discuss the result with him.

Use Hadoop, Hive, and Spark and focus on the following detail. Explore any cloud solutions are available for this use case and make any significant difference to the current architecture.

System Architecture – Identify the required Infrastructure. Analyze and find out the necessary configurations required for the dev environments and for production.

Network Architecture - Install and connect all hardware, software, and firmware as per logical and physical design within the schedule, budget, and quality of the proposed project, all labels required for every interface, connector, links, node, and all devices listed in the equipment and budget list.

Algorithm and Mathematical model - The algorithm describes the operation of the system.
<u>System Operation</u>
Building a model to detect credit card fraudulent transactions. Use Kafka to stream data to Hive table and export the file to CSV. Export the data to notebooks and clean the data.

Data cleaning will include eliminating unnecessary fields and dealing with outliers. To perform supervised machine learning projects, you need to have labelled data. Labelled data are datasets with labels or information for classification, in this case, there will be the transaction and its label specifying if it is a fraudulent or genuine transaction. Based on this label or categories a model is to be built which will recognize the pattern and the same model will be used to predict if the transaction is fraudulent or not. Test data will be used to evaluate the performance of the model built. Before evaluating the performance in the test data set, you can use the training data set to make the validation

Mathematical Models

Use python scripts for creating a different model. This is a classification problem, where the transaction is to be categorized as whether they are fraudulent or not. You can try using different classification algorithms and see how they perform. Some of the classification algorithms are

- Logistics Regression
- Naïve Bayes
- K-Nearest Neighbors
- Decision Tree
- Support Vector Machines

Result And Evaluation

Use confusion matrix and other measures such as accuracy and F1 score as evaluation metrics.

Implementation

Implementation of the project includes the use of Kafka, Hadoop, and Hive for data streaming and collection purposes and notebooks and python scripts for analytical purposes.

Deliverables for This Week's Case Study

- Make sure all the required applications are installed and configured to implement the use case.
- Prepare an architecture diagram for the used Hadoop ecosystem and analytics steps.
- Identify the different libraries of python to be used for data cleaning, transformation, and modelling processes.
- Research the classification algorithm and its working mechanism.
- Build a model using an ideal algorithm for this scenario and rely on F1 score and accuracy to choose the model.

Week 9

Applicable VLOs or EESs for This Week's Case Study

3. Design and apply data models that meet the needs of a specific operational process or business model.

5. Collaborate effectively with diverse teams to design and present data visualizations that communicate Big Data concepts and information to stakeholders and business professionals.

This Week's Detailed Case Study Information

Data scientists implement exploratory data analysis tools and techniques to investigate, analyze, and summarize the main characteristics of datasets, often utilizing data visualization methodologies. EDA techniques allow for effective manipulation of data sources, enabling data scientists to find the answers they need by discovering data patterns, spotting anomalies, checking assumptions, or testing a hypothesis.

Deepsight Analytics has a Power BI Premium subscription for its data visualization needs. This week you are collaborating with Francis to perform EDA on the cleaned data. In earlier weeks you cleaned the data set, now it's time to explore more about the data set and find anomalies and outliers. You will be using both python libraries and Power BI to visualize the data and learn more about it.

This step is a continuation of a model-building process that you started in earlier weeks. Data visualization helps to analyze data, and find hidden patterns, anomalies and outliers.

- Use different plots and graphs to visualize the data
- Compare the visualization built using python libraries and power BI.
- Note down the facts or notions in the dataset you were able to see because of the visualization
- Build the ML model again with the insight you received after EDA with visualization.

Deliverables for This Week's Case Study

- Perform Exploratory Data Analysis (EDA) on the data and build the model using a classification algorithm.
- List the insights about data you were able to see due to data visualizations.
- Build a model after doing EDA with visualizations and compare the models.
- Discuss the use case scenarios for Data visualizations using python libraries and Power BI.

Week 10

Applicable VLOs or EESs for This Week's Case Study

3. Design and apply data models that meet the needs of a specific operational process or business model.

4. Develop software applications, algorithms, and artificial intelligence models to manipulate, correlate and reduce data sets and produce project documentation and reports.

9. Deliver data-oriented projects using data science, business analysis, and project management principles, tools, and techniques.

This Week's Detailed Case Study Information

Data analytics is an iterative process. Analyst spending more time exploring the data and building models will develop more knowledge of the data. Collaboration with Data scientists and domain experts will help to gain a unique perspective on data insights. You have developed a model using an algorithm to classify the transactions in weeks 8 and 9 Use different algorithms and compare the performance of models. The use of an algorithm depends on the use case, not all algorithms will be applicable to all classification problems. You will be discussing why you chose a particular algorithm over other algorithms with Francis, so you should get your fact and supporting points ready.

Deepsight analytics is planning to integrate the developed model into a web application using REST API. Every time a transaction happens the system would send that transaction details to the application and identify if the transaction is fraudulent or not. Claire has asked you to come up with the IT infrastructure to implement the project in web application. Conduct quick research to find out other ways in which the models can be used to detect fraudulent transactions in real-time.

- Research the cross-fold validation and evaluate your model using the same.
- Use different folds for k-cross-fold validation and analyze the result.
- Research how the built models can be saved and reused.

Deliverables for This Week's Case Study

- List the pros, cons, and trade-offs of using your choice of algorithm.
- Make a report highlighting the performance of different models and explaining how cross validation assist in measuring algorithm performance.
- Research the ways to implement these models in real time transaction and predict the possible outcomes.

Week 11

Applicable VLOs or EESs for This Week's Case Study

3. Design and apply data models that meet the needs of a specific operational process or business model.

4. Develop software applications, algorithms, and artificial intelligence models to manipulate, correlate and reduce data sets and produce project documentation and reports.

9. Deliver data-oriented projects using data science, business analysis, and project management principles, tools, and techniques.

This Week's Detailed Case Study Information

Once teams move from a stage where they are occasionally updating a single model to having multiple frequently updating models in production, a pipeline approach becomes paramount. A machine learning pipeline helps in automating machine learning workflows. Machine learning pipelines are iterative as every step is repeated to continuously improve the accuracy of the model and achieve a successful algorithm. The code is split into more manageable components such as data validation, model training, model evaluation, and re-training triggering.

In the case of normal machine learning workflow, the model is the product and in automated workflows, the pipeline is the product. Even an ad hoc model can be deployed in real time depending on the use case. Save the ad hoc model (without pipeline) using pickle and Joblib libraries. In earlier weeks you have tried to build an ad hoc model and the code might not be reusable. Try to build a pipeline for the same workflow, this time making the reusable code. Refer to Sklearn documentation to make pipelines. To make the work reusable and implement the modular design in ML workflow create pipelines.

- Research about the ML pipelines and the benefits of using pipelines
- Create functions for every step that you build in ML workflow in earlier weeks
- Create a pipeline for the whole process
- Compare this modular approach with the normal steps that you followed in previous weeks

Deliverables for This Week's Case Study

- Submit both the pickle and joblib file.
- Submit the python scripts created for the pipeline implementation.

Week 12

Applicable VLOs or EESs for This Week's Case Study

8. Implement data security solutions in compliance with corporate security policies, ethical standards, and industry regulations.

This Week's Detailed Case Study Information

Data security is always one of the major concerns when it comes to Big Data or any Data Science projects. There is always the possibility of data theft, attacks on systems and ransomware. Credit card data theft or breach may provide personal information such as an address, spending habits, email address, and phone numbers which might result in phishing emails and calls and might further damage too.

In the previous weeks, you applied some data masking techniques to hide the information. Claire has asked you to make sure that all PII should be hidden so that the data cannot be related to any individual. Your team for this project is responsible to provide security at data level, Deepsight analytics has a different team working for maintaining cloud and network security. They are mostly responsible for maintaining the security of data in cloud storage.

GridSearchCV is the process of performing hyperparameter tuning in order to determine the optimal values for a given model. The performance of a model significantly depends on the value of hyperparameters. Note that there is no way to know in advance the best values for hyperparameters so ideally, to know the optimal values all possible values of hyperparameters should be tried. Doing this manually could take a considerable amount of time and resources and thus GridSearchCV can be used to automate the tuning of hyperparameters. Claire has asked you to identify ideal Hyperparameters using GridSearchCV. This week, you need to

- Identify the security challenges
- Evaluate the risks and take precautions.
- Research on GridSearchCv and tune hyperparameters using GridSearchCV.

Deliverables for This Week's Case Study

1. In this week, your team needs to submit a report including:
   - What kind of attack can an attacker use to stole credit card information.
   - What can be the consequences of big data theft and strategies to prevent them.
2. Find out the ideal algorithm and hyperparameters to be used for the best results.

Program Name: Big Data Analytics

Project Code: **CPL-5559-DSMM**

Week 13

Applicable VLOs or EES for This Week's Case Study

7. Identify and assess data analytics and Big Data business strategies and workflows to respond to new opportunities or provide project solutions.

9. Deliver data-oriented projects using data science, business analysis, and project management principles, tools, and techniques.

This Week's Detailed Case Study Information

Submission of Project Report + Practice Presentation

- Finish Project Report for submission your final submission is due next week. Be proud of the work you have completed in this project, now you can spend time polishing your presentation and making sure you will capture the stakeholder's attention in a positive way.
- Review APA Guidelines and ensure your project has followed them. This is particularly important.

Hone your presentation skills.

- A Presentation for your Fraudulent Transaction detection project is meant to highlight your research findings and the conclusions, opportunities, and best practices that you can be followed on other projects. The analysis of your findings is one of the most important parts and should be conveyed in your presentation.

Deliverables for This Week's Case Study

1. Final Project Report – this is your final document with all supporting resources: including any appendices. Bibliography and reference in APA format required.
2. Feedback Video
   - Prepare to answer questions regarding the project on client expectations, Job Market, and on how you will sell your product

Week 14

Preparing for Your Final Week Activities

It is the end of your work term. Your supervisor is grateful for your efforts. The final week contains activities which include both individual and teamwork efforts. Take this opportunity to shine bright in the final activities.

Final Week Deliverables and Format Requirements

Your supervisor will provide you with more detail about the Final Week responsibilities.

AIP Project Completion

Following completion of the Final Week activities, you will be notified by your supervisor if you pass or fail the WIL Project.

Appendix

Acronym Used

HDFS: Hadoop Distributed File system
PCA: Principal Component Analysis
PD: Personal Development
HQL: Hive query Language
SQL: Structured Query Language
RBAC: Role Based Access Control
ML: Machine Learning
DE: Data Engineer
PM: Project Manager
PII: Personal identifiable information
AI: Artificial Intelligence