WIL Project Case Study Orientation

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

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

Week 1

This Week's Detailed Case Study Information

Congratulations you have been hired as a Data Architect for 'Stride' located in Vaughn ON.

Stride is a leading logistics company that specializes in fast and reliable delivery services for businesses and individuals. With a network of local and international partners, Stride is able to offer a wide range of shipping options to meet the needs of its customers. Whether you need to ship a small package across town or a large pallet of goods across the country, Stride has a solution for you. The company's state-of-the-art logistics system allows them to track shipments in real-time and ensure that packages are delivered on time, every time.

In addition to traditional shipping services, Stride also offers a range of value-added services such as warehousing, distribution, and customs clearance. This allows businesses to outsource their logistics needs and focus on what they do best. Stride is dedicated to providing the highest level of customer service and satisfaction. Their team of experienced professionals is always ready to help with any questions or concerns, and they are committed to finding the best solutions for their clients. Whether it's a small business owner or a global corporation, Stride has the resources and expertise to handle all logistic needs.

The analytics department at Stride is a vital part of the company's operations, responsible for providing insights and data-driven decision-making to support the company's growth and success. The team at Stride's analytics department is composed of highly skilled data scientists and analysts who use advanced analytics tools and techniques to analyze and interpret large amounts of data from various sources, including customer data, shipping data, and financial data.

One of the key responsibilities of the analytics department is to use this data to identify trends and patterns that can inform strategic business decisions. For example, they might analyze shipping data to identify inefficiencies in the company's operations or use customer data to identify opportunities for growth. The analytics department also plays a key role in supporting the development of new products and services. By analyzing customer data, they can help identify new product ideas or areas for improvement in existing products.

Overall, the analytics department at Stride is a critical component of the company's success, using data and analytics to drive informed decision-making and drive business growth.

The company has stored the data for every shipment, and they want to analyze this data now to answer the following questions:

- Was the product delivered on the expected time and what was the customer rating?
- Was the customer's query answered

- If Product importance is high. having the highest rating or being delivered on time?

With this question answered, Stride can focus on increasing the efficiency of their work and find the ideal method of doing their work. This analytics will significantly help them to succeed in the work they are doing. You will be working with the following team as Data Architect:

- Diamond Beasley (Lead Data Architect)
- Holden Booker (Machine Learning Engineer)
- Myah Dunn (Data and Analytics Manager)
- Lisa Diaz (Project Coordinator)
- Krish Burns (Data Scientist)

Diamond is a highly experienced data architect at Stride, with a strong background in designing and implementing data storage and management systems. He is known for his attention to detail and ability to think critically about complex problems and is an asset to the team with his deep understanding of data structures and architectures. Holden is a talented ML engineer at Stride, with a passion for using machine learning to solve real-world problems. He is known for his creativity and ability to think outside the box and is always looking for ways to apply ML techniques in innovative ways. Myah is the Data Analytics manager at Stride, with a wealth of experience in leading data-driven projects and teams. She is known for her strong analytical skills and ability to translate complex data insights into actionable business recommendations. Lisa is a skilled project coordinator at Stride, with a background in project management and customer service. She is known for her ability to keep projects on track and for her excellent communication skills, which make her a valuable asset to the team. Krish is a data scientist at Stride, with a strong background in data analysis and machine learning. He is known for his analytical mind and ability to identify patterns in large datasets and is always looking for ways to use data to drive business growth.

Overall, the team at Stride is composed of highly experienced and skilled professionals, each with their own unique strengths and personalities. Together, they bring a diverse set of skills and expertise to the table and are able to work effectively as a team to drive business success.

Week 1 Onboarding Expectations and Participation

Your task this week is to participate in training and orientation for Telewire 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 possible to highlight your talent as a group. You would be presenting these deliverables to Tess 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

- Collect, house, extract, manipulate, maintain and mine data sets that respond to organizational, financial, or market needs.
- Recommend different systems and network architectures, artificial intelligence, and data storage technologies to support data analytics and Big Data.

This Week's Detailed Case Study Information

This week you will be working with Diamond to collect the data in order to facilitate further analysis processes.

Stride has been using the Hadoop ecosystem to store their data, and as part of the data collection process for this project, you will need to export the data from Hadoop before performing analysis on it. One way to do this is by using Sqoop, a tool that allows you to transfer data between Hadoop and structured data stores such as relational databases. To use Sqoop to get the data from Hadoop to your local system, you will need to connect to the Hadoop cluster and specify the location of the data that you want to export. You can then use Sqoop to extract the data and save it to a local file in CSV format.

Before starting the data export process, it is important to ensure that you have the necessary permissions to access the data and that you have the correct version of Sqoop installed on your system. You should also make sure that you have enough storage space on your local system to accommodate the data that you are exporting. Once, you have exported the data from Hadoop using Sqoop, you can begin the process of analyzing and manipulating the data as needed for your project. This might involve cleaning and transforming the data, or applying machine learning algorithms to extract insights and make predictions.

Deliverables for This Week's Case Study

1. Install Hadoop and it ecosystem with a single cluster.
2. Connect to the Hadoop cluster and verify access to the data that needs to be exported.
3. Install and configure Sqoop on the local system, if necessary.
4. Use Sqoop to extract the data from Hadoop and save it to a local file in CSV format.
5. Review the project description and scope to understand the overall goals and deliverables of the project.

Week 3

Applicable VLOs or EESs for This Week's Case Study

- 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

Now the data collection process has been completed before you do any further work you need to plan the overall project. Before beginning the data analysis portion of the project, it is important to plan out the overall project and create a work breakdown structure (WBS). A WBS is a visual representation of the project that shows the different tasks and deliverables that are required to complete the project. It is a useful tool for project managers because it helps to break down the project into smaller, manageable pieces and allows them to visualize the overall project scope. This week, you will need to develop a WBS for the project. This will involve outlining all the steps of the project work, including any tasks or deliverables that are required. The final project deliverable should be at the top of the WBS, with the tasks and work packages needed to complete it listed underneath. The WBS should be hierarchical, with the higher levels representing the broad project scope and the lower levels representing more specific tasks and work packages.

Since the data collection process has already been completed, you can skip the data migration step in the WBS. However, you should still include any tasks or deliverables that are related to preparing the data for analysis, such as cleaning or transforming the data as needed. By creating a comprehensive WBS, you will be better prepared to manage the project and ensure that all the necessary work is completed on time.

On the basis of the description of project provided on week 1, Lisa has asked you to create a WBS documents

Deliverables for This Week's Case Study

1. Outline the final project deliverable at the top of the WBS.
2. Break down the project scope into smaller tasks and work packages, organizing them in a hierarchical structure.
3. Identify any dependencies or resources that are required for each task or work package. Estimate the time and resources needed for each task or work package.
4. Review the WBS with the project team to ensure that all necessary tasks and deliverables are included.
5. Verify that the data has been successfully exported and is in the correct format.

Week 4

Applicable VLOs or EESs for This Week's Case Study

- 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

Before starting a new project, it is important to identify the input, tools and technologies, and output (ITTO) that will be involved. In this project, the input is the datasets provided by Stride, the tools and technologies will be chosen by the team, and the output will be a prediction model.

When selecting the tools and technologies for a project, it is important to consider the skillset of the team and the ideal stack for solving the problem at hand. It is also important to stay up to date on emerging technologies and their alternatives, as well as their limitations. As an ML engineer, it is important to have a strong understanding of the different technologies available and how they can be used to solve different problems.

In addition to technical skills, it is also important to have domain knowledge in the specific application or field that the project is focused on. In this case, it would be helpful to gather knowledge about the data lifecycle in the logistics industry and the specific challenges of data analysis in this domain.

To ensure success in the project, it can be helpful to research the tools, services, and software components that will be required, as well as review projects that have been done in similar areas and gather recommendations from those projects. It is also important to list out the specific challenges that may be encountered in the data analysis process in the logistics industry. By gathering this information, the team can be better prepared to tackle the project and achieve their desired outcomes.

Deliverables for This Week's Case Study

1. Identify specific features for the model and any additional data sources needed.
2. Assess the quality and relevance of data sources and determine if data cleaning or transformation is necessary.
3. Set up a data storage solution for the project.
4. Implement version control and create a project folder structure.
5. Develop a project plan with timeline and milestones for data preparation, model development, and evaluation.

Week 5

Applicable VLOs or EESs for This Week's Case Study

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

This Week's Detailed Case Study Information

This week the team will be reviewing the data and cloud security approaches taken in this project. You will be collaborating with Krish, Lisa and Myah to maintain the highest level of security in this domain.

Lisa wants to analyze the risks involved in these projects since there is customer information in the given dataset, she is being extra careful about the security of the data and the possibility of a data breach. She has asked the team to identify if there is any personal identifiable information (PII) and mask them before moving to further analysis.

Krish suggested that the team implement a hierarchy of roles and permissions for the project to ensure that members have access to the resources they need while minimizing the risk of accidental or intentional alterations to the resources. As part of this process, it will be your responsibility to analyze the different roles and permissions required for the project and come up with a combination that meets the needs of the team. This may involve considering the different tasks that team members will be responsible for, the resources they will need to access, and any potential risks that need to be mitigated. It's important to strike a balance between giving team members the access they need to do their jobs effectively and protecting the resources from unauthorized or accidental changes.

Lisa has requested you create a secure way to access the services in the cloud, you will research possible ways to prevent data breaches and implement a few of them. As part of maintaining Cloud security, you will identify the role and permissions needed for those roles for the different users in the project.

Deliverables for This Week's Case Study

1. List down the roles and associated responsibilities needed by each team member for this project.
2. Review the roles and responsibilities, and create the permissions required to perform those activities.
3. Check if the permission you've created is there by default in the cloud services of your choice, if not create custom roles.

Week 6- Mid Term Week

Mid-Term Panel Evaluation Preparation:

The team will prepare for the Mid-Term Panel Evaluation this week. For the Team Presentation create a professional multimedia presentation highlighting the key aspects of your project thus far. Please see Moodle for full details.

Presentation and Oral Delivery

CONTENT

- Overview of work in the Iconic
- Highlight three key areas you find of interest:
  - Two areas related to weekly work completed
  - One area to highlight PD or other activity
- Apply reflecting skills
- Present the importance/benefit of work to Iconic.

WIL PROJECT

Program Name: Big Data Analytics

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

Week 7

Applicable VLOs or EESs for This Week's Case Study

- 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

The team has successfully extracted and loaded the data from the identified data sources into the data storage solution. They have also performed some basic data exploration and visualization to get a better understanding of the key trends and patterns in the data.

Now, Myah has asked you to focus on developing the data pipelines that will be used to extract, transform, and load the data on an ongoing basis. This is an important step, as the team will need to ensure that the data is cleaned and transformed consistently and efficiently in order to support the Shipment prediction model. In addition to working on the data pipelines, Holden is also setting up a data testing and quality assurance process to ensure the integrity and accuracy of the data. This involves writing test cases and developing automated scripts to validate the data at various stages of the data pipeline.

Finally, you are beginning to work with Krish on the model development tasks. You will be selecting and testing different algorithms, tuning model parameters, and evaluating the model's performance. By the end of this week, Lisa hopes to have made significant progress on these tasks

Deliverables for This Week's Case Study

1. Identify the specific data sources and data storage solutions that will be used in the project.
2. Develop data pipelines to extract, transform, and load the data on an ongoing basis.
3. Set up a data testing and quality assurance process to ensure the integrity and accuracy of the data.
4. Write test cases and develop automated scripts to validate the data at various stages of the data pipeline.
5. Export the data to notebooks and perform preliminary analysis. The preliminary analysis would include the following steps.
    - Verify the units of measurement
    - Clean" the data by deleting or, if possible, correcting obviously incorrect results.
    - Identify outliers or otherwise unusual results.
    - Check for missing data

WIL PROJECT

Program Name: Big Data Analytics

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

Week 8

Applicable VLOs or EESs for This Week's Case Study

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

The data given to you contains features stating if the shipment was delivered on time or not. This is a target variable, thus the provided data is labelled data. This data is historic data and on the basis of this data, you will try to predict the behaviour of every transaction. With all features given, you would try to predict if the shipment will be delivered on time or not. Krish is impressed by the job you have done so far but, building a data model is not an easy task it requires knowledge of the different domains. So try your best to prove to him that you are competent in model building as well. As a step in ML workflow create a model to predict if the shipment will be delivered in time or not.

At this point, all data are migrated to the Azure SQL database, and the ML workflow should be started. There are multiple options for building the infrastructure for ML, either you can rely on the notebook instance provided by Azure or any other cloud services or perform it locally on your machine but it is suggested to use cloud services. At this point, you have performed a preliminary analysis, and are ready to the start model-building process. There are multiple options for building the infrastructure for ML, either you can rely on the notebook instance provided by cloud services or perform it locally on your machine. When the data set is small your local machine might suffice but as the data grows or when you start using extensive models and neural networks the computing power in the local machine might not be enough so it is suggested to use cloud services.

Lisa has asked you to calculate the cost of running the operation in the cloud and on-premises machine and present the report comparing the cost. Explore any cloud solutions that are available for this use case and make any suggestions for optimal solutions. It's always wise to use graphs and charts to compare budgets and resources in reports.

Deliverables for This Week's Case Study

1. Build a model using an ideal algorithm for this scenario and rely on the F1 score and accuracy to choose the model
2. Identify the different libraries of python to be used for data cleaning, transformation, and modelling processes.
3. Select and test different algorithms to be used in the model.

WIL PROJECT

Program Name: Big Data Analytics

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

Week 9

Applicable VLOs or EESs for This Week's Case Study

- 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

Krish has a unique way of getting more information and hidden pattern out of the data, he visualizes the data in a preliminary analysis, get insights out of it and relying on this knowledge builds a model after analyzing the model he performs exploratory analysis again. He believes that when you build a model you get to know more about your data and doing one more iteration of data EDA really helps to unravel the hidden pattern in data.

Exploratory data analysis (EDA) is an important step in understanding and preparing a dataset for analysis. It involves using various techniques, such as Pandas, NumPy, statistical methods, and data visualization, to examine the characteristics and relationships between different elements of the dataset. Some benefits of EDA include:

- Improved understanding of the dataset and identification of issues such as missing values or human error
- Insight into the key features of the dataset and their relationships
- Identification of important variables and removal of unnecessary variables
- Detection of outliers or unusual data points

Overall, EDA helps to ensure that the dataset is cleaned and ready for further analysis, and provides valuable insights into the key features and relationships within the data. In this week, you will work with Krish to perform EDA.

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.

Deliverables for This Week's Case Study

1. Perform Exploratory Data Analysis (EDA) on the data and build the model using a classification algorithm.
2. Use different plots and graphs to visualize the data.
3. Can python libraries be used instead of Tableau in every use case?
4. Compare the visuals made using python Libraries and Tableau and highlight the differences and similarities of using these two different tools.

WIL PROJECT

Program Name: Big Data Analytics

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

Week 10

Applicable VLOs or EESs for This Week's Case Study

- 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

Krish is aware of the importance of properly tuning hyperparameters in order to achieve the best performance from a machine learning model. He has asked you to tune the hyperparameters for the model using GridSearchCV, a method that combines grid search with cross-validation.

Hyperparameters are variables that are specified by the user when building a machine-learning model. They are used to evaluate the optimal parameters of the model and their values are chosen by the model builder. In this case, you will be using GridSearchCV to tune the hyperparameters.

GridSearchCV involves dividing the data into train and test sets and further dividing the train set into train and validation sets. The model is then trained on the train set and evaluated on the validation set. This process is repeated for different combinations of hyperparameter values, and the combination that yields the best performance is chosen.

In this week you will be working to tune the hyperparameters using GridSearchCV. Rest of the portion of ML workflow will remain the same. Relying upon the same preliminary analysis and feature engineering that you did in earlier weeks you will perform hyperParameter tuning.

Deliverables for This Week's Case Study

1. Use the following four models and range of features to find out the performance of each model:
    - Logistic Regression
    - Decision Tree
    - XGBoost
    - SVM

2. List the hyperparameters in which the algorithm performed well and find out the model with the best performance.
3. Explain the possibility of overfitting and underfitting of model and different techniques to handle imbalance dataset.

WIL PROJECT

Program Name: Big Data Analytics

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

Week 11

Applicable VLOs or EESs for This Week's Case Study

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

This Week's Detailed Case Study Information

Neural networks, also known as artificial neural networks or simulated neural networks, are a type of machine learning algorithm that is based on the structure of the human brain. They are particularly useful for modelling complex, nonlinear relationships between input and output data, and can be used to make intelligent decisions with limited human intervention.

In the past, you have used traditional machine learning methods to develop models. Lisa has asked you to try using a neural network to see if it can improve the performance of the model. While there is no guarantee that a neural network will lead to an improvement in performance or accuracy, it is worth trying as it may provide a deeper understanding of the data. In data science, there is no single correct answer, and it is important to continuously try to improve the performance of the model by exploring different approaches.

Some commonly used Neural Networks are
- Deep Neural Network
- Wide Neural Network
- Deep and wide neural Networks

Before diving into a Neural network, it is wise to revise your knowledge of it. Neural Networks can get complex pretty fast so, go through the following steps to refresh your knowledge in Neural Networks

- Go through the documentation of TensorFlow and Keras.
- List out the use cases for different types of Neural network.
- Create a table comparing the performance of the conventional machine learning model and Neural Networks. And conclude if it makes sense to use NN in this use case

Deliverables for This Week's Case Study

1. Use Deep neural networks and wide neural networks to classify the activities in data.
2. Compare the performance of the two models.
3. Change the number of nodes in hidden layers and add some more hidden layers and compare the performance with the rest of the models.
4. Increase the epochs for model and compare the loss in different models.

WIL PROJECT

Program Name: Big Data Analytics

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

Week 12

Applicable VLOs or EESs for This Week's Case Study

- 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

A machine-learning pipeline is a series of steps that allow data to flow into and out of a machine-learning model or model. This includes raw data input, features, outputs, the model itself and its parameters, and prediction outputs. The design and implementation of a machine learning pipeline are critical to the performance and effectiveness of enterprise AI software applications. It is also important to consider factors such as the choice of machine learning libraries and runtime environments, such as processor requirements, memory, and storage.

Object-oriented code can be useful in the context of a machine-learning pipeline because it allows for reusable and extendable code. This means that the same functions and classes can be used in multiple analyses or projects, reducing the time it takes to produce results. Additionally, object-oriented code can be easier to debug because it is typically organized in a clear and logical way.

In some cases, a machine learning model may be used as a standalone product, while in other cases, the pipeline itself is the product. It is also possible to deploy an ad hoc model in real time for certain use cases. To save an ad hoc model (without a pipeline), libraries such as pickle and Joblib can be used.

Lisa has asked you to build a pipeline for the same workflow, this time making the reusable code. Refer to Sklearn documentation to make pipelines.

Deliverables for This Week's Case Study

1. Create an object-oriented approach for ML workflow created in earlier weeks
2. Create a pipeline for the whole machine learning process and create separate python files for each process
3. Compare this modular approach with the normal steps that you followed in previous weeks
4. Submit both the pickle and joblib file.
5. Submit the python scripts created for the pipeline implementation.

WIL PROJECT

Program Name: Big Data Analytics

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

Week 13

Applicable VLOs or EES for This Week's Case Study

- Identify and assess data analytics and Big Data business strategies and workflows to respond to new opportunities or provide project solutions.
- 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 this 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 Sentiment analysis 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

WIL PROJECT

Program Name: Big Data Analytics

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

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.

WIL 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
NLP: Natural Language processing
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: Personally identifiable information
AI: Artificial Intelligence
ITTO: Input Tools& Technologies and output.