**FULL STACK DATA SCIENTIST AND NLP PRACTITIONER**

Passionate about mathematics and analytical fields, with a strong affinity for working with numbers and data. He has several years of experience in machine learning (mostly in Natural Language Processing space), data engineering, and data visualization, and he possesses strong problem-solving skills, cross-functional collaboration abilities, and a commitment to transforming complex data into actionable insights.

Self-learner and self-motivated; energetic, enthusiastic and driven; resilience under pressure; curious & eager to learn; adaptable to cultural differences; able to build a relationship with stakeholder.

**HIGHLIGHTS AND CERTIFICATION**

## **3rd Place in 2015 Hilti IT Innovation Competition - Big Data Analytics**

An international big data analytics competition organized by Hilti Malaysia. Content of the competition is to propose a recommender system for Hilti Online (HOL) website which selling construction tools. Applied Alternating Least Square (ALS) algorithm under Apache Spark platform for developing the prototype of recommender system.

1. **Rapid Miner Certified Analyst**

Took the Rapid Miner Analyst course on April 2017 and passed the exam on June 2017.

1. **Microsoft Certified: Azure Fundamentals**

Took the course on Feb 2020 and passed the exam on Feb 2020.

1. **Able to self-mange and leading other**

Demonstrates strong self-management skills and experienced in leading teams by setting clear objectives, mentoring colleagues, and fostering a supportive team dynamic.

# **PROJECTS URL**

1. Simple RAG Application - <https://huggingface.co/spaces/warless/RAG_application>

**Description** : A prototype RAG (Retrieval-Augmented Generation) application which allows users to perform simple NLP tasks on their own file.

1. News Article Classification Application - <https://huggingface.co/spaces/warless/Article_Classification>

**Description** : An application that allows users to paste a news article, and returns the article's category while flagging identified entities. (news source : <https://www.nst.com.my/>)

# **TECHNICAL SKILLS**

1. **Programming/Scripting Language** : Python, SQL, PySpark , R, PowerQuery and DAX.
2. **Machine Learning** : Supervised and unsupervised machine learning algorithm , NLP (sentiment analysis , article classification ,entity recognition and etc) and deep learning.
3. **Data Engineering** : Azure Databricks, Azure Data Factory, Azure Synapse and Microsoft Fabric.
4. **Data Visualization** : PowerBI, Tableau, Python Package (Dash, Seaborn, Plotly).
5. **Database** : Azure SQL (MSQL), MySQL, PostgreSQL.
6. **Cloud Services** : Microsoft Azure , Google Cloud Platform(BigQuery and App Engine).
7. **API and Framework** : Python Flask, Streamlit , Git, Docker , ElasticSearch.
8. **Others** : Web crawling/scraping , Power Automate.

# **WORK EXPERIENCE**

## **British American Tobacco (8h May 2023 – Present)**

### **MANAGER – DATA SCIENCE**

1. Optimize business process of SOx (Sarbanes-Oxley Act) , CN (Control Navigator) and other business control related compliance process by leveraging data analytics and machine learning.

* **PowerAutomate** : To automate process which could integrate with PowerAutomate.
* **PowerBI** : Reporting automation and insights visualization
* **Microsoft Fabric** : Leverage different services in Microsoft Fabric to build an end-to-end data pipeline projects.

1. Gather requirements from various stakeholders and oversee and manage a project team consisting of both external and internal members in the development and delivery of the project after building the prototype/MVP.

## **DKSH Corporate Shared Services Center (13th July 2020 – 7th May 2023)**

### **TEAM LEAD– DATA SCIENCE**

Leverage Microsoft Azure Services to build a full stack data science pipeline

* **Azure Databricks** : Platform to develop/deploy data science scripts which written in Python , PySpark or R.
* **Azure Blob Storage** : Object Storage to stored temporary input/output file.
* **Azure SQL/Azure Synapse** : Data warehouse or database to store input data and output data from data science processes.
* **Azure Data Factory** : Orchestration tool to automate, schedule and monitor Azure services which interconnected in data science pipeline.

Develop project for new use cases or manage certain on-going projects.

* Agile approach on developing and supporting on-going projects which used by different countries.
* Communicate and gather requirement with regional users when rolling out existing product to new country.
* Develop POC or prototype for new use cases initiated by initial users and productionize them on Azure Cloud after users satisfied on the final version.

Leading data science production on Azure Cloud in the team.

* Lead and help on every data science project production on Azure Cloud.
* Giving guidance to team members on related services on Azure Cloud.
* Optimization and modification on performance of code written in Azure Databricks.

## **MALAYSIA AIRLINES (15th May 2019 – 10th July 2020)**

### **SENIOR DATA SCIENTIST**

Leverage Microsoft Stack and Qubole Cloud Solution to deliver daily reporting job.

* Used Microsoft Power Bi desktop and web application to extract data from Qubole database and develop reports to other team (Digital Marketing, Payment, Ancillary Sales and etc..)
* Automate certain reporting job using Microsoft Flow (now being called Microsoft Power Automate).
* Utilize some of the Microsoft Azure API to support certain advance analytic for example Natural Language Processing job.
* Schedule reporting script written in SQL, Python or R on Qubole Environment and store end results in Microsoft Azure blob storage and directory.

Utilize Open Source Package to perform advance modelling.

* Perform text analytic and sentiment analysis using R and Python open source package.
* Develop traveller type prediction using Python open source Logistic Regression package.
* Built travel destination recommendation engine using PySpark ALS (Alternating Least Square) open source package.

Support Ad-Hoc Request from other team and manage certain on-going projects

* Extract data from intended period for other team to analyse campaign performance, payment data etc.
* Support projects from other team which required the expertise from data analytic team by involve in their recurring meeting, understand business purpose, help to extract certain insights to verify business assumptions etc.

## **Ciro Solution (2nd May 2018 – 14th May 2019)**

### **BIG DATA ANALYST**

Building ETL Pipeline and automate reports.

* Setting up big data lake using Hadoop stack (Spark, Hive and HBase).
* ETL tools included Spark (Scala and PySpark ), Apache Kafka, Hadoop , Hive and etc.
* Reports included player summary, game summary, platform summary etc.
* Automate reporting pipeline using Airflow as prove of concept.
* Stream data from MySQL database to Hive using Kafka Confluent.

Proposed potential analysis and data model for data analytic purpose.

* Develop RFM model for game and player:
* Game - Analyse game performance based on day since last login, amount of players login each day, and amount of money put in the slotting game.
* Player - Group player into different groups based on their day since last login to game platform, number of times they login and play the games and amount money put in the slotting game.

Shift production of ETL Platform to Google Cloud

* Utilize Google Cloud components such as Google BigQuery, AppEngine and Cron to develop daily reporting platform for reports such as player daily summary, player RFM , game RFM reports and etc.
* The functions and module which required to build the reports needed are written in Python (Pandas, BIgquery SQL and etc.) and converted to Flask application. The Flask application is then deployed to AppEngine and schedule using Google Cron job.

## **Media Prima Digital (2rd January 2017 – 30th Apr 2018)**

### **DATA SCIENCE EXECUTIVE**

1. Recommendation System Prototype

* Extracted viewership data such as movie title, number of views through API provided by video analytic portal and perform Market Basket Analysis on the datasets extracted. (using R)
* Approach changed to Cosine Similarity using Python.

1. Report Automation

* Viewership, transaction and subscription data from video analytic portal side frequently injected to our data warehouse by data engineers.
* Queried data using SQL and put them into Tableau worksheet and dashboard to visualize the weekly video analytic reports
* This report automation help to save about 4 man days of manual (power point and excel) reporting task.

1. Online Advertisement Campaign Analysis

* One of the campaign analysis is the Nissan car campaign advertisement which been posted to three of Media Prima NSTP portal sides (New Straits Times, Berita Harian and Harian Metro)
* The analysis involved determine behaviour and pattern of users respond on the advertisement posted before they submitted their personal details.
* This analysis helps to understand the performance of this online advertisement campaign in NSTP side.

1. Automate Article Category tagging and Keyword

* Used Python API package to extract the articles body, category from the three NSTP print side and perform the article classification training and testing.
* Utilized the Python NLP package ‘Spacy’ to perform keyword (entity) extraction.
* Model was deployed as restful API using Python Flask and deploy to Heroku app as production.

1. Automate Keyword tagging for advertisement audience targeting

* Automated the keyword tagging process for keywords coming from NSTP side before put into Lotame system for online audience targeting.
* For example, ‘Jack Ma’ should auto-tagged under ‘Business’ category.
* This automation helps to save human manual tagging time.

## **Hilti Asia IT Services Sdn Bhd (3rd Aug 2015 – 31st Dec 2016)**

### **SAP BYDESIGN IT PROCES CONSULTANT**

* Managed, compiled, uploaded and downloaded web reports, excel reports in Hilti SAP ByDesign system.
* Supported tickets and incidents which raised by end users regarding reporting issue in SAP ByDesign system.
* Interacted and communicated with business end users and SAP service provider.
* Designed new web reports/ excel reports in Hilti SAP ByDesign system based on requirement from business end users from other countries.

## **Fusionex Corp International (3rd Feb 2014 – 31st Jul 2014)**

## **DATA SCIENCE INTERNSHIP**

* Performed basic analytic on retail business data sets Sunway, Mydin, Sam's Grocery and others.
* Performed predictive analysis and classification on retail business data sets.
* Performed data virtualization on results obtained from analytic.
* Web crawling/scraping on social media sites like Facebook and Twitter and some business blogs to extract significant comments for further analytic for example sentiment analysis.
* Performed simple Natural Language Processing (NLP) task.
* Tools used are R and Python.

# **EDUCATION**

Bachelor Degree of Computer Science in University Sains Malaysia(Sep 2011 — Jun 2015)

* Major in Distributed System and Security & Minor in Mathematics.