Course Outline

Section 1 - Data - Driven Business

Module 1.1

Introduction to Applied Data Science from Hitachi Vantara -
3 hrs ILT or VILT

- Course overview
- What is Hitachi Vantara?
- Evidence based / data-driven decision-making
- Introduction to data science
- What's in it for me
- Exercise – research into "data-driven"

Module 1.2

Your Opportunity
Be a driver of success in data-driven initiatives - 2 hrs ILT or VILT

- Why so many data-driven projects fail
- Economic value vs technological innovation
- What difference can you make?
- Exercise – personal value proposition and goals

Module 1.3

The importance of Big Data and IoT (from the "Big Data MBA", Bill Schmarzo) - 1.5 hrs WBT

- Big Data Business Model Maturity Index
- What is Big Data and why is it so important
- Basic data science concepts
- Creating Smart Spaces
- Simplifying advanced analytics
- Monetizing your IoT
- Determining economic value of data
- Five laws of digital transformation

Module 1.4

Thinking Like a Data Scientist (TLADS) Methodology - 5 hrs WBT/ Book

- TLADS Methodology Overview
- The Art of Thinking Like a Data Scientist (book by Bill Schmarzo)

Module 1.5

- • Interactive, scenario based workshop – TLADS methodology leading to completion of Hypothesis Canvas.

Hypothesis Development Canvas Workshop (TLADS) - 4 hrs ILT or VILT

Section 2 - Exploring Data Science

Module 2.1

Introduction to data science and statistics techniques - 6 hrs ILT or VILT

- The process of problem solving
- Descriptive statistics – mean, median, mode, standard deviation, variance, inferential statistics Sample, population, regression, hypothesis testing, correlation, co-variance
- Data collection and visualization

Module 2.2

Data analysis methods -
10 hrs ILT or VILT

- Overview of data analysis methods, Performing Exploratory Data Analysis, Feature Exploration, Types of Variables, Univariate Analysis, Multivariate Analysis
- The Normal Distribution, Handling Missing Data Applying QC method, Outlier Treatment Applying multivariate analysis
- Application of Probability Theory and Statistical Hypothesis Testing, Business application

Module 2.3

Ethical AI - 3 hrs WBT

- Ethical AI and why it matters
- Exercise – research and write article

Module 2.4

Introduction to Artificial Intelligence and Machine Learning - 2 hrs ILT or VILT

- Introduction to AI, Evolution & Revolution of AI, Introduction of Applications in various Domains (Scientific including Health Science, Engineering, Financial services and other industries)

Module 2.5

Industrial IoT Primer - 2 hrs WBT

- Introduction to IoT
- Two worlds – IT and OT
- OT terminology
- Industrial process control
- Spotlight on manufacturing
- Vendor landscape
- Digital Transformation of OT

Module 2.6

Introduction to Python (can skip if prior experience on Python) –
12 hrs ILT or VILT

- Introduction to Python, The Python Interpreter, Working with Command Line/ IDLE Python Data Types, Built in operators, Functions and Methods, Block and Indentation

	• Conditional Statements, Control Flow Statements, Data Structures • The Zip Function , Range, Collection framework, Membership and Identity operators, Functions , Default Arguments, Lambda Expressions, Map Filter and Reduce Functions, Scope of Variables • Object Oriented Programming, Instantiating & Inheritance Classes • Method and Operator Over Loading
Module 2.7	
Python for Machine Learning - 2 days ILT or VILT	• Handing the Exceptions, File I/O Handling, Regular Expressions, Serializing Python Objects, Django: Web Development Programming, Flask: Web Development Programming • Data Science Programming Tools, Pandas Introduction, Constructing Series Objects, Constructing Data Frame Objects, Reading CSV Files into Data Frame Object Reading Excel Files into Data Frame Object, Selecting columns and Slicing, Row and Column Filtering, Finding Unique values in DataFrame, Delete Duplicates in Data Frame, Merging and Concatenating DataFrames, Pivoting, Group by Operations, Working with Dates • Introduction to Numpy, Comparison on Memory and run-time with native list, Function to Create Numpy Arrays, Attributes of Arrays, Indexing of Arrays, Vectorized operations • Introduction to Matplotlib, Plotting Line Plot, Plotting with Categorical Data, Plotting Scatter Plots • Using Seaborn to Visualize Data, Density plots, Histograms, Heat Maps, Violin Plots
Module 2.8	
Introduction to Machine Learning - 2 days ILT or VILT	• Overview of machine learning, What is machine learning? • Types of machine learning, Process of machine learning, Examples of • Applying machine learning, machine learning and data mining, Machine learning and deep learning • Verifying machine learning with analysis tool, Scikit Learn, H2O Framework

	• Sample Selection, Training Data, Testing Data, Validation Data Feature Scaling, Standardization • Linear Regression, Multiple Linear Regression Logistic Regression, Multi Logistic Regression Gradient Boosting Algorithm • Model Validation, Confusion Matrix, ROC Curve, Cross Validation AUC, R2 Value, Lift, Gain, K-fold Validation • Bootstrapping & Bagging, Over Fitting vs Under-fitting Diagnosis • SMOTE, Random Over Sampling, Random Under Sampling • Probabilistic Classifier - Naive Bayes Classifier, Non-probabilistic • Classifier - K-nearest neighbor (KNN), Decision Tree, Random Forest Algorithm, Support Vector Machines (SVM) • Need For Dimensionality Reduction, Principal Component Analysis (PCA) • Hierarchical and K-means Clustering • Case 1: Optimization • Case 2: Anomaly detection • Case 3: Numerical prediction • Implementing machine learning, Exercise - four lab exercises
Section 3 - Data Operations	
Module 3.1	
Introduction to data operations - 2 hr WBT	• Right data, right place, right time • The changing landscape • Agile, DataOps and DevOps • Containerization • GIGO • Governance • Data catalogues
Module 3.2	
Analytics Data Pipeline - 2 hrs ILT or VILT	• Traditional data pipelines – OLTP, OLAP, ETL • Big Data challenges and the need for a new way • Tools, vendor ecosystem • Modern Data Lakes • Analytics maturity levels - DEPPA • IOT and Edge Computing • Data sources • Exercise – data pipeline vendor map

Module 3.3	
Big Data Ecosystem - 2 hrs ILT or VILT	• Hadoop ecosystem (HDFS, MapReduce, Spark etc) • NoSQL • Streaming data • Public Cloud
Module 3.4	
Data Modelling and Visualization - 3 hrs ILT or VILT	• Introduction to data modelling and visualization • Common visualization techniques • Exercise – view some visualization examples using Pentaho and Excel
Module 3.5	
Project example using Lumada - 3 hr WBT	• Use case outline • Business initiative and KPI's • Features / Deliverables in waves • Focus on value • Solution overview • Key learnings • Exercise – research an industry domain
Module 3.6 (STEM)	
Pentaho Data Integration - 1 day ILT or VILT	• Introduction to Pentaho Data Integration • Extract, Transform and Load • Transformations and Jobs • Making data operations more Agile (metadata injection, machine learning orchestration, self-service) • Exercise – my first transformation and job
Module 3.7 (STEM)	
Pentaho Platform with Business Analytics - 1 day ILT or VILT	• Introduction to the Pentaho Platform • Traditional business intelligence (charts, dashboards etc) • Embedding – powering applications with analytics • Data modelling basics (Mondrian Cubes) • Exercise – build a data model and visualization with Pentaho • Exercise – various hands-on exercises throughout
Module 3.8 (STEM)	
Streaming Data with Pentaho (IoT) - 2 hrs WBT	• MQTT • Kafka • Kinesis

Section 4 - Applied Data Science Project	
Module 4.1	
Project write-up 8+ hrs exercise	• Builds on activities throughout the course • Pulls concepts together and applies them to a scenario
Module 4.2	
Final workshop 4 hrs ILT or VILT	• Sharing of ideas and next steps • Final wrap-up of the course

*ILT = Instructor Led Training, these sections will be delivered by face to face or virtually by an instructor.