**Module Synopsis**

**Specialist Diploma in Data Science (Data Engineering)**

**PDC 1: PDC in Fundamentals of Data Science (120 hours)**

**MS9001: Introduction to Statistics for Data Science (60 hours)**

This module provides students with an introduction to elementary probability theory and statistical concepts and principles that lay the foundation to understand and learn the statistical procedures and methods in the subsequent modules. The topics covered include descriptive statistics, rules of probability, probability distributions of discrete and continuous random variables, sampling distributions, statistical estimation, and hypothesis testing.

**IT8701: Introduction to Programming for Data Science (60 hours)**

This module provides students with the fundamental skills to code applications to retrieve, clean and visualize data using the Python programming language. Students learn key concepts such as what structured and unstructured data are, and how they can create and manipulate relational and NoSQL databases to explore data and to create visualizations that can help them gain useful insights from it.

**PDC 2: PDC in in Data Engineering (150 hours)**

**IT8704: Data Management Systems (75 hours)**

This module provides participants to learn about database concepts, architectural design for building data warehousing solutions. The topics covered include Data concept, OLTP vs OLAP, data validation techniques and data integration concepts.

**IT8705: Big Data Technologies (75 hours)**

This module provides participants with an understanding of what Big Data is, why it is needed and how Big Data pipelines are implemented. It also covers Apache Spark Libraries. Apache Spark is a powerful platform that provides users with new ways to store and make use of big data. Students will analyse data using Spark SQL, explore running machine learning algorithms using Mlib and demonstrate how to create a streaming analytics application using Spark Streaming.