

Module Title : **Certified Data Science Specialist**

Duration : **5 days**

Prerequisites

All participants should have basic understanding of data, relations, and basic knowledge of mathematics.

Who Should Attend

This workshop is intended for individuals who are interested in learning data science, or who want to begin their career as a data scientist.

Exam Format

The CDSS Certification Exam duration is 2 hours, consisting of 50 Multiple Choice Questions, with a Passing Score of 70%. You will receive a professional CDSS Certification upon passing the exam.

Course Overview

Our lives are flooded by large amounts of information, but not all of them are useful data. Therefore it is essential for us to learn how to apply data science to every aspect of our daily life from personal finances, reading and lifestyle habits, to making informed business decisions. In this course you will learn how to leverage on data to ease life, or unlock new economic value for a business. This course is a hands-on guided course for you to learn the concepts, tools, and techniques that you need to begin learning data science. We will cover the key topics from data science to big data, and the processes of gathering, cleaning and handling data. This course has a good balance of theory and practical applications, and key concepts are taught using case study references. Upon completion, participants will be able to perform basic data handling tasks, collect and analyze data, and present them using industry standard tools.

Learning Outcomes

Upon completion of this course, you will be able to:

- Identify the appropriate model for different data types
- Create your own data process and analysis workflow
- Define and explain the key concepts and models relevant to data science.
- Differentiate key data ETL process, from cleaning, processing to visualization.
- Implement algorithms to extract information from dataset.
- Apply best practices in data science, and become familiar with standard tools.

Course Outline

Day 1

Introduction to Data Science

- What is Data?
- Types of Data?
- What is Data Science?
- Statistical thinking
- Knowledge Check
- Lab activity

Data Process

- Extract, Transform and Load (ETL)
- Data Cleansing
- Aggregation, Filtering, Sorting, Joining
- Data Workflow
- Knowledge Check
- Lab activity

Data Quality

- Raw vs Tidy Data
- Key features of data quality
- Maintenance of data quality
- Data profiling
- Data completeness and consistency

Life of a data scientist

- Identify problem
- Define question
- Define ideal dataset
- Obtain data
- Analyze data
- Interpret results
- Distribute results
- Knowledge Check

Day 2

Beginning Database

- Types of Databases
- Relational Databases
- NoSQL
- Hybrid database
- Knowledge Check
- Lab activity

Structured Query Language (SQL)

- Performing CRUD (Create, Retrieve, Update, Delete)
- Designing a Real world database
- Normalizing a table
- Knowledge Check
- Lab activity

Introduction to Python

- Basics of Python language
- Functions and packages
- Python lists
- Functional programming in Python
- Numpy and Scipy
- iPython
- Knowledge Check
- Lab activity

Lab: Exploring data using Python

Day 3

Data Gathering

- Obtain data from online repositories
- Import data from local file formats (json, xml)
- Import data using Web API
- Scrape website for data
- Knowledge Check
- Lab activity
- Instructor-led case study

Exploratory Data Analysis

- What is EDA?

- Goals of EDA
- The role of graphics
- Handling outliers
- Dimension reduction

Introduction to R

- Features of R
- Vectors
- Matrices and Arrays
- Data Frame
- Input / Output

Lab: Exploring data using R

Day 4

Introduction Text Mining

- What is Text Mining?
- Natural Language Processing
- Pre-processing text data
- Extracting features from documents
- Using BeautifulSoup
- Measuring document similarity
- Knowledge Check
- Lab activity

Supervised Learning

- What is prediction?
- Sampling, training set, testing set
- Constructing a decision tree
- Knowledge Check
- Lab activity

Day 5

Presenting Data

- Choosing the right visualization
- Plotting data using Python libraries
- Plotting data using R

- Lab activity
- Using Jupyter Notebook to validate scripts
- Knowledge Check
- Lab activity

Data Analysis Presentation

- Using Markdown language
- Convert your data into slides
- Data presentation techniques
- The pitfall of data analysis
- Knowledge Check
- Lab activity
- Group presentation

Lab: Mini Project

Big Data Landscape

- What is small data?
- What is big data?
- Big Data Analytics vs Data Science
- Key elements in Big Data (3Vs)
- Extracting values from big data
- Challenge in Big Data

Big Data Tools and Applications

- Introduction Hadoop Ecosystem
- Cloudera vs Hortonworks
- Real world big data applications
- Knowledge Check
- Group discussion

What's Next?

- Preview of Data Science Specialist
- Showing advanced data analysis techniques
- Demo: Interactive visualizations