**Module Title : Cloudera Data Scientist Training**
**Duration : 4 days**

## Overview

This four-day workshop covers data science and machine learning workflows at scale using Apache Spark 2 and other key components of the Hadoop ecosystem. The workshop emphasizes the use of data science and machine learning methods to address real-world business challenges.

Using scenarios and datasets from a fictional technology company, students discover insights to support critical business decisions and develop data products to transform the business. The material is presented through a sequence of brief lectures, interactive demonstrations, extensive hands-on exercises, and discussions. The Apache Spark demonstrations and exercises are conducted in Python (with PySpark) and R (with sparklyr) using the Cloudera Data Science Workbench (CDSW) environment.

## What to Expect

The workshop is designed for data scientists who currently use Python or R to work with smaller datasets on a single machine and who need to scale up their analyses and machine learning models to large datasets on distributed clusters. Data engineers and developers with some knowledge of data science and machine learning may also find this workshop useful.

Workshop participants should have a basic understanding of Python or R and some experience exploring and analyzing data and developing statistical or machine learning models. Knowledge of Hadoop or Spark is not required.

## Topics: Cloudera Data Scientist Training

The workshop includes brief lectures, interactive demonstrations, hands-on exercises, and discussions covering topics including:

- Overview of data science and machine learning at scale
- Overview of the Hadoop ecosystem
- Working with HDFS data and Hive tables using Hue
- Introduction to Cloudera Data Science Workbench
- Overview of Apache Spark 2
- Reading and writing data
- Inspecting data quality

- Cleansing and transforming data
- Summarizing and grouping data
- Combining, splitting, and reshaping data
- Exploring data
- Configuring, monitoring, and troubleshooting Spark applications
- Overview of machine learning in Spark MLlib
- Extracting, transforming, and selecting features
- Building and evaluating regression models
- Building and evaluating classification models
- Building and evaluating clustering models
- Cross-validating models and tuning hyperparameters
- Building machine learning pipelines
- Deploying machine learning models

## Technologies

Participants gain practical skills and hands-on experience with data science tools including:

- Spark, Spark SQL, and Spark MLlib
- PySpark and sparklyr
- Cloudera Data Science Workbench (CDSW)
- Hue