Cloudera Data Science Workbench

MASTER DATA SCIENCE

DATA SCIENCE | MACHINE LEARNING | BIGDATA | DEEP LEARNING | CLOUD SERVICES

cloubia

Cloubia Technologies Pvt. Ltd. Bangalore
+91-7877962524 | www.cloubia.com | info@cloubia.com
Bigdata Hadoop, Spark, Machine Learning, Deep Learning, Cloud AWS & Google Cloud, Data Science

COURSE DURATION

Why this Course?

- Businesses Will Need One Million Data Scientists by 2018 - KDnuggets
- Roles like chief data & chief analytics officers have emerged to ensure that analytical insights drive business strategies - Forbes
- The average salary for a Data Scientist is $113k (Glassdoor)

- **100+ Mentoring Hours**
- **2 Hours Per Session**
- **5+ Projects**
- **20+ Case Studies**

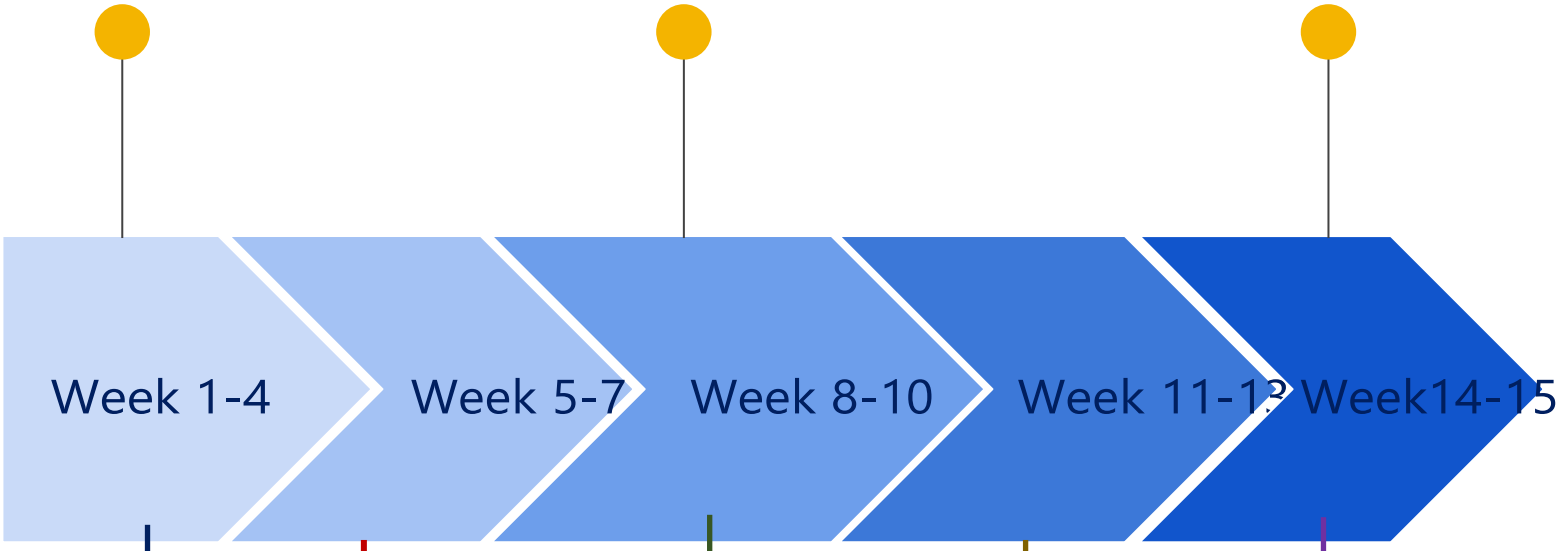
COURSE DURATION

100+ hours,
Live online sessions
By Industry experts to
Become industry ready
Data Scientist

Machine Learning with Python
Foundations of statistics, regressions, classifications, model selections, unsupervised learning, time series analysis, NLP, deep learning, Tensorflow, etc.

Machine learning theory defense, Capstone project presentations.

Code reviews, resume workshop, mock interviews, career day



1. Basic Linux System,
2. Data Scientist Tools (Git),
3. Google or AWS Cloud
4. Hadoop Ecosystem,
5. Pig,
6. Hive

1. Spark using Python
2. RDD,
3. Spark-SQL,
4. Graphx,
5. Streaming,
6. Bigquery/ Redshift

1. Statistics
2. Data Analysis
3. Visualization
4. Tableau

Machine Learning
Scikit
Spark-ML
Supervised
Unsupervised
Clustering
Classification
Recommendation System

Deep Learning
Artificial Intelligence
Natural Language Processing
Computer Vision
Tensorflow

COURSE SPECIFICATIONS

Course Objectives

After the completion of the course, you should be able to:

- Gain insight into the 'Roles' played by a Data Scientist
- Analyze several types of data using python
- Describe the Data Science Life Cycle
- Work with different data formats like XML, CSV etc.
- Learn tools and techniques for Data Transformation
- Discuss Data Mining techniques and their implementation
- Analyze data using Machine Learning algorithms in python
- Explain Time Series and it's related concepts
- Perform Text Mining and Sentimental analyses on text data
- Gain insight into Data Visualization and Optimization techniques
- Understand the concepts of Deep Learning

Why Learn Data Science?

The incorporation of technology in our everyday lives has been made possible by the availability of data in enormous amounts. Data is drawn from different sectors and platforms including cell phones, social media, e-commerce sites, various surveys, internet searches, etc.

However, the interpretation of vast amounts of unstructured data for effective decision making may prove too complex and time consuming for companies, hence, the emergence of Data Science.

Data science incorporates tools from multi disciplines to gather a data set, process and derive insights from the data set, extract meaningful data from the set, and interpret it for decision-making purposes.

The disciplinary areas that make up the data science field include mining, statistics, machine learning, analytics, and some programming. Data mining applies algorithms in the complex data set to reveal patterns which are then used to extract useable and relevant data from the set.

Statistical measures like predictive analytics utilize this extracted data to gauge events that are likely to happen in the future based on what the data shows happened in the past. Machine learning is an artificial intelligence tool that processes mass quantities of data that a human would be unable to process in a lifetime.

Machine learning perfects the decision model presented under predictive analytics by matching the likelihood of an event happening to what actually happened at the predicted time.

Who should go for this Course?

The course is designed for all those who want to learn about the life cycle of Data Science, which would include acquisition of data from various sources, data wrangling and data visualization. Applying Machine Learning techniques in R language, and wish to apply these techniques on different types of Data.
The following professionals can go for this course:
1. Developers aspiring to be a 'Data Scientist'
2. Analytics Managers who are leading a team of analysts
3. Business Analysts who want to understand Machine Learning (ML) Techniques
4. Information Architects who want to gain expertise in Predictive Analytics
5. 'R' professionals who want to captivate and analyze Big Data
7. Analysts wanting to understand Data Science methodologies

What are the pre-requisites for this Course?

There is no specific pre-requisite for the course, however basic understanding of R can be beneficial. Cloubia offers you a complimentary self-paced course, i.e. "R Essentials" when you enroll in Data Science Certification Training.

What are the system requirements for this course?

If you have a Windows system you should have:

- Microsoft Windows 7 or newer (32-bit and 64-bit)
- Microsoft Server 2008 R2 or newer
- Intel Pentium 4 or AMD Opteron processor or newer
- 2 GB memory
- 1.5 GB minimum free disk space
- 1366 x 768 screen resolution or higher

If you have a MAC system you should have :

- iMac/MacBook computers 2009 or newer
- OSX 10.10 or newer
- 5 GB minimum free disk space
- 1366 x 768 screen resolution or higher

COURSE CURRICULUM

Section-01 Linux Basics

Installing Ubuntu in virtualbox

ipaddress,

hostname,

ssh,

package installation,

java installation,

.bashrc file,

tar file,

scp commands

Basic essential linux commands

Advance essential linux commands

User Group,

File Permissions Management

Section-02 Bigdata Hadoop Ecosystem

Bigdata Introduction and HDFS File System

Hadoop Installation on Single Node and Multinode

Cloudera CDH Virtual Machine Introduction

Cloudera Manager Multinode Cluster Installation on Google Cloud

Cloudera Manager Multinode Cluster Installation on AWS

HDFS Commands

Sqoop

Flume

PIG

Hive

HBase

Oozie

Mapreduce

Section-03 Python

Python – I

- Installation of Python and Ipython Notebook
- Python Objects
- Number & Booleans
- Strings

Python – II

- Container objects
- Mutability of objects
- Operators – Arithmetic & Bitwise
- comparison and Assignment operators
- Operators Precedence and associativity.

Python – III

- Conditions (If else if-elif-else)
- Loops (While, for)
- Break and Continue statements
- Range functions

Python – IV

- String object basics
- String methods
- Splitting and Joining Strings
- String format functions
- list object basics
- list methods
- List as stack and Queues
- List comprehensions

Python – V

- Tuples, Sets, Dictionary Object basics
- Dictionary Object methods
- Dictionary View Objects.
- Functions basics, Parameter passing
- Iterators, Lambda functions
- Map, Reduce, filter functions

Python – VI

- OOPS basic concepts
- Creating classes and Objects
- Inheritance
- Multiple Inheritance
- Working with files
- Reading and writing files

Python – VII

- Using Standard Module
- Creating new modules
- Exceptions Handling with Try-except
- Creating
- inserting and retrieving Table
- Updating and deleting the data.

Python – VIII

Numpy

Python – IX

Scipy

Python – X

Data Loading, Storage, and File Formats

Section-04 Spark

Spark-I Introduction

Spark-II Installation

Spark-III Spark Integration with Different ID's

Spark-IV RDD

Spark-V Read data from different sources

Spark-VI Spark Debugging and tuning

Spark-VII Reading different type of data

Spark-VIII Google Bigquery

Spark-IX Amazon Redshift

Spark-X Spark-SQL

Spark-XI Spark Streaming

Spark- XII Spark Graphx

Spark-XIII Kafka

Spark-XIV Submit Spark jobs on Google Cloud Dataproc

Spark- XV Submit Spark jobs on Amazon AWS

Section-05 Data Analysis

Data Analysis- I Getting Started with pandas

Data Analysis- II Data Wrangling: Clean, Transform,Merge, Reshape

Data Analysis- III Plotting and Visualization

Data Analysis- IV Data Aggregation and Group Operations

Data Analysis- V Matplotlib

Data Analysis- VI Tableau

Section-06 Statistics

Statistics-1

- Descriptive Statistics
- Sample vs Population statistics
- Random Variables
- Probability distribution function
- Expected value

Statistics-II

- Binomial Distribution
- Normal Distributions
- z-score
- Central limit Theorem

Statistics-III

- Hypothesis testing
- Z-Stats vs T-stats
- Type 1 type 2 error
- confidence interval

Statistics-IV

- Chi Square test

Statistics-V

- ANOVA test
- F-stats

Section-07 Machine Learning (Lecture 111-150)

Machine Learning - 1

- Introduction
- Supervised
- Unsupervised
- Semi-supervised
- Reinforcement
- Train
- Test
- Validation Split
- Performance
- Overfitting
- Underfitting

Machine Learning - II

- Linear Regression
- Assumptions
- R square adjusted R square
- Intro to Scikit learn
- Training methodology
- Hands on linear regression

Machine Learning - III

- Logistics regression
- Precision Recall
- ROC curve
- F-Score

Machine Learning - IV

- Decision Tree
- Cross Validation
- Bias vs Variance

Machine Learning - V

- Ensemble approach
- Bagging Boosting
- Randon Forest
- Variable Importance

Machine Learning - VI

- XGBoost
- Hands on XgBoost

Machine Learning - VII

- K Nearest Neighbour
- Lazy learners
- Curse of Dimensionality
- KNN Issues

Machine Learning - VIII

- Text Ananlytics
- Tokenizing
- Chunking
- Document term Matrix
- TFIDF
- Sentiment analysis hands on

Machine Learning - IX

- Hierarchical clustering
- K-Means
- Performance measurement

Machine Learning - X

- Principal Component analysis
- Dimesionality reduction
- Factor Analysis

Machine Learning - XI

- Time Series Forecasting
- Moving Average
- ARIMA model

Section-08 Deep Learning

Deep Learning - I

- Basic of Neural Network
- Type of NN
- Cost Function
- Gradient descent
- Linear Algebra basics

Deep Learning - II

- Vanilla implementation of Neural Network in python

Deep Learning - III

- Tensorflow basics
- Hands on Simple NN with Tensorflow

Deep Learning - IV

- Word Embedding
- CBOW
- Skip-gram
- Word Relations
- Hands on word2vec

Deep Learning - V

- Convolutional Neural Network
- Maxpool
- Window padding
- Hands On

Deep Learning - VI

- Image classification using Convolutional Neural Network

Deep Learning - VII

- Recurrent Neural Network
- Long Short-Term Memory (LSTM) architecture
- Building Story writer using character level RNN

Deep Learning - VIII

- Sentiment Analysis Hands on
- Hands on embedding + RNN

Deep Learning - IX

- Seq-to-Seq model
- Hands on translation

Deep Learning - X

- Encoder Decoder
- Hands on cleaning images

Deep Learning - XI

- GAN
- Generative Model Using GAN

Deep Learning - XII

- Semi-supervised learning using GAN

Deep Learning - XIII

- Restricted Boltzmann Machine(RBM) and Autoencoders

Section-09 Natural Processing Language

Neural Networks and Deep Learning

Improving Deep Neural Networks

Speech API Using Google Cloud

Translation API Using Google Cloud

Structuring Machine Learning Projects

Convolutional Neural Networks

Sequence Models

Section-10 Computer Vision

TensorFlow-Keras Introduction

Setting Up TensorFlow Environment

TensorFlow- Keras Loss Functions

TensorFlow-Keras Evaluation Metrics

TensorFlow-Keras Optimizers

CNNs with TensorFlow-Keras

Vision API Using Google Cloud

TensorFlow- Keras Layers

TensorFlow-Keras Functional API

Image Preprocessing and Augmentation

Image Classification with VGG

Cat and Dog Dataset

VGG Network Architecture

VGG Implementation in TensorFlow-Keras

Model Training and Evaluation

Transfer Learning – Feature Extraction

Transfer Learning - Fine Tuning