edureka!
Discover Learning

PYTHON SPARK CERTIFICATION TRAINING USING PYSPARK

Course Curriculum : Your 13 module Learning Plan

https://www.edureka.co/pyspark-certification-training

About Edureka

Edureka is a leading e-learning platform providing live instructor-led interactive online training. We cater to professionals and students across the globe in categories like Big Data & Hadoop, Business Analytics, NoSQL Databases, Java & Mobile Technologies, System Engineering, Project Management and Programming. We have an easy and affordable learning solution that is accessible to millions of learners. With our students spread across countries like the US, India, UK, Canada, Singapore, Australia, Middle East, Brazil and many others, we have built a community of over 1 million learners across the globe.

About Course

Edureka's PySpark Certification Training is designed to provide you the knowledge and skills that are required to become a successful Spark Developer using Python and prepare you for the Cloudera Hadoop and Spark Developer Certification Exam (CCA175). Throughout the PySpark Training, you will get an in-depth knowledge of Apache Spark and the Spark Ecosystem, which includes Spark RDD, Spark SQL, Spark MLlib and Spark Streaming. You will also get comprehensive knowledge of Python Programming language, HDFS, Sqoop, Flume, Spark GraphX and Messaging System such as Kafka.

Python Spark Course Curriculum

Introduction to Big Data Hadoop and Spark

Learning Objectives: In this module, you will understand Big Data, the limitations of the existing solutions for Big Data problem, how Hadoop solves the Big Data problem, Hadoop ecosystem components, Hadoop Architecture, HDFS, Rack Awareness, and Replication. You will learn about the Hadoop Cluster Architecture, important configuration files in a Hadoop Cluster. You will also get an introduction to Spark, why it is used and understanding of the difference between batch processing and real-time processing.

Topics:

- What is Big Data?
- Big Data Customer Scenarios
- Limitations and Solutions of Existing Data Analytics Architecture with Uber Use Case
- How Hadoop Solves the Big Data Problem?
- What is Hadoop?
- Hadoop's Key Characteristics
- Hadoop Ecosystem and HDFS
- Hadoop Core Components
- Rack Awareness and Block Replication
- YARN and its Advantage
- Hadoop Cluster and its Architecture
- Hadoop: Different Cluster Modes
- Big Data Analytics with Batch & Real-Time Processing
- Why Spark is Needed?
- What is Spark?
- How Spark Differs from its Competitors?
- Spark at eBay

- Spark's Place in Hadoop Ecosystem

Introduction to Python for Apache Spark

Learning Objectives: In this module, you will learn basics of Python programming and learn different types of sequence structures, related operations and their usage. You will also learn diverse ways of opening, reading, and writing to files.

Topics:

- Overview of Python
- Different Applications where Python is Used
- Values, Types, Variables
- Operands and Expressions
- Conditional Statements
- Loops
- Command Line Arguments
- Writing to the Screen
- Python files I/O Functions
- Numbers
- Strings and related operations
- Tuples and related operations
- Lists and related operations
- Dictionaries and related operations
- Sets and related operations

Hands-On:

- Creating "Hello World" code
- Demonstrating Conditional Statements

- Demonstrating Loops
- Tuple - properties, related operations, compared with list
- List - properties, related operations
- Dictionary - properties, related operations
- Set - properties, related operations

Functions, OOPs, and Modules in Python

Learning Objectives: In this Module, you will learn how to create generic python scripts, how to address errors/exceptions in code and finally how to extract/filter content using regex.

Topics:

- Functions
- Function Parameters
- Global Variables
- Variable Scope and Returning Values
- Lambda Functions
- Object-Oriented Concepts
- Standard Libraries
- Modules Used in Python
- The Import Statements
- Module Search Path
- Package Installation Ways

Hands-On:

- Functions - Syntax, Arguments, Keyword Arguments, Return Values
- Lambda - Features, Syntax, Options, Compared with the Functions
- Sorting - Sequences, Dictionaries, Limitations of Sorting

- Errors and Exceptions - Types of Issues, Remediation
- Packages and Module - Modules, Import Options, sys Path

Deep Dive into Apache Spark Framework

Learning Objectives: In this module, you will understand Apache Spark in depth and you will be learning about various Spark components, you will be creating and running various spark applications. At the end, youwill learn how to perform data ingestion using Sqoop.

Topics:

- Spark Components & its Architecture
- Spark Deployment Modes
- Introduction to PySpark Shell
- Submitting PySpark Job
- Spark Web UI
- Writing your first PySpark Job Using Jupyter Notebook
- Data Ingestion using Sqoop

Hands-On:

- Building and Running Spark Application
- Spark Application Web UI
- Understanding different Spark Properties

Playing with Spark RDDs

Learning Objectives: In this module, you will learn about Spark - RDDs and other RDD related manipulations for implementing business logics (Transformations, Actions, and Functions performed on RDD).

Topics:

- Challenges in Existing Computing Methods
- Probable Solution & How RDD Solves the Problem
- What is RDD, It's Operations, Transformations & Actions
- Data Loading and Saving Through RDDs
- Key-Value Pair RDDs
- Other Pair RDDs, Two Pair RDDs
- RDD Lineage
- RDD Persistence
- WordCount Program Using RDD Concepts
- RDD Partitioning & How it Helps Achieve Parallelization
- Passing Functions to Spark

Hands-On:

- Loading data in RDDs
- Saving data through RDDs
- RDD Transformations
- RDD Actions and Functions
- RDD Partitions
- WordCount through RDDs

DataFrames and Spark SQL

Learning Objectives: In this module, you will learn about SparkSQL which is used to process structured data with SQL queries. You will learn about data-frames and datasets in Spark SQL along with different kind of SQL operations performed on the data-frames. You will also learn about the Spark and Hive integration.

Topics:

- Need for Spark SQL

- What is Spark SQL
- Spark SQL Architecture
- SQL Context in Spark SQL
- Schema RDDs
- User Defined Functions
- Data Frames & Datasets
- Interoperating with RDDs
- JSON and Parquet File Formats
- Loading Data through Different Sources
- Spark-Hive Integration

Hands-On:

- Spark SQL – Creating data frames
- Loading and transforming data through different sources
- Stock Market Analysis
- Spark-Hive Integration

Machine Learning using Spark MLlib

Learning Objectives: In this module, you will learn about why machine learning is needed, different Machine Learning techniques/algorithms and their implementation using Spark MLlib.

Topics:

- Why Machine Learning
- What is Machine Learning
- Where Machine Learning is used
- Face Detection: USE CASE
- Different Types of Machine Learning Techniques

- Introduction to MLlib
- Features of MLlib and MLlib Tools
- Various ML algorithms supported by MLlib

Deep Dive into Spark MLlib

Learning Objectives: In this module, you will be implementing various algorithms supported by MLlib such as Linear Regression, Decision Tree, Random Forest and many more.

Topics:

- Supervised Learning: Linear Regression, Logistic Regression, Decision Tree, Random Forest
- Unsupervised Learning: K-Means Clustering & How It Works with MLlib
- Analysis of US Election Data using MLlib (K-Means)

Hands-On:

- K- Means Clustering
- Linear Regression
- Logistic Regression
- Decision Tree
- Random Forest

Understanding Apache Kafka and Apache Flume

Learning Objectives: In this module, you will understand Kafka and Kafka Architecture. Afterwards you will go through the details of Kafka Cluster and you will also learn how to configure different types of Kafka Cluster. After that you will see how messages are produced and consumed using Kafka API's in Java. You will also get an introduction to Apache Flume, its basic architecture and how it is integrated with Apache Kafka for event processing. You will learn how to ingest streaming data using flume.

Topics:

- Need for Kafka
- What is Kafka
- Core Concepts of Kafka
- Kafka Architecture
- Where is Kafka Used
- Understanding the Components of Kafka Cluster
- Configuring Kafka Cluster
- Kafka Producer and Consumer Java API
- Need of Apache Flume
- What is Apache Flume
- Basic Flume Architecture
- Flume Sources
- Flume Sinks
- Flume Channels
- Flume Configuration
- Integrating Apache Flume and Apache Kafka

Hands-On:

- Configuring Single Node Single Broker Cluster
- Configuring Single Node Multi-Broker Cluster
- Producing and consuming messages through Kafka Java API
- Flume Commands
- Setting up Flume Agent
- Streaming Twitter Data into HDFS

Apache Spark Streaming - Processing Multiple Batches

Learning Objectives: In this module, you will work on Spark streaming which is used to build scalable fault-tolerant streaming applications. You will learn about DStreams and various Transformations performed on the streaming data. You will get to know about commonly used streaming operators such as Sliding Window Operators and Stateful Operators.

Topics:

- Drawbacks in Existing Computing Methods
- Why Streaming is Necessary
- What is Spark Streaming
- Spark Streaming Features
- Spark Streaming Workflow
- How Uber Uses Streaming Data
- Streaming Context & DStreams
- Transformations on DStreams
- Describe Windowed Operators and Why it is Useful
- Important Windowed Operators
- Slice, Window and ReduceByWindow Operators
- Stateful Operators

Hands-On:

- WordCount Program using Spark Streaming

Apache Spark Streaming - Data Sources

Learning Objectives: In this module, you will learn about the different streaming data sources such as Kafka and flume. At the end of the module, you will be able to create a spark streaming application.

Topics:

- Apache Spark Streaming: Data Sources
- Streaming Data Source Overview
- Apache Flume and Apache Kafka Data Sources
- Example: Using a Kafka Direct Data Source

Hands-On:

- Various Spark Streaming Data Sources

Implementing an End-to-End Project

Project 1- Domain: Finance
Statement: A leading financial bank is trying to broaden the financial inclusion for the unbanked population by providing a positive and safe borrowing experience. In order to make sure this underserved population has a positive loan experience, it makes use of a variety of alternative data--including telco and transactional information--to predict their clients' repayment abilities. The bank has asked you to develop a solution to ensure that clients capable of repayment are not rejected and that loans are given with a principal, maturity, and repayment calendar that will empower their clients to be successful.

Project 2- Domain: Media and Entertainment
Statement: Analyze and deduce the best performing movies based on the customer feedback and review. Use two different API's (Spark RDD and Spark DataFrame) on datasets to find the best ranking movies.

Spark GraphX (Self-Paced)

Learning Objective: In this module, you will be learning the key concepts of Spark GraphX programming concepts and operations along with different GraphX algorithms and their implementations.

Topics:

- Introduction to Spark GraphX
- Information about a Graph

- GraphX Basic APIs and Operations
- Spark GraphX Algorithm - PageRank, Personalized PageRank, Triangle Count, Shortest Paths, Connected Components, Strongly Connected Components, Label Propagation

Hands-On:

- The Traveling Salesman problem
- Minimum Spanning Trees

Python Spark Training Projects

How will I execute the practicals in this PySpark Certification Training?

You will execute all your PySpark Course Assignments/Case Studies in the Cloud LAB environment provided by Edureka. You will be accessing the Cloud LAB via a browser. In case of any doubt, Edureka's Support Team will be available 24*7 for prompt assistance.

What are the system requirements PySpark Training Course?

You don't have to worry about the system requirements as you will be executing your practicals on a Cloud LAB which is a pre-configured environment. This environment already contains all the necessary tools and services required for Edureka's PySpark Training.

Which projects and case studies will be a part this Edureka's PySpark Online Training Course?

At the end of the PySpark Training, you will be assigned with real-life use-cases as certification projects to further hone your skills and prepare you for the various Spark Developer Roles. Following are few industry-specific case studies that are included in our Apache Spark Developer Certification Training.

Project 1- Domain: Financial
Statement: A leading financial bank is trying to broaden the financial inclusion for the unbanked population by providing a positive and safe borrowing experience. In order to make sure this underserved population has a positive loan experience, it makes use of a variety of alternative data--including telco and transactional information--to predict their clients' repayment abilities. The bank has asked you to develop a solution to ensure that clients capable of repayment are not rejected and that loans are given with a principal, maturity, and repayment calendar that will empower their clients to be successful.

Project 2- Domain: Transportation Industry
Business challenge/requirement: With the spike in pollution levels and the fuel prices, many Bicycle Sharing Programs are running around the world. Bicycle sharing systems are a means of renting

bicycles where the process of obtaining membership, rental and bike return is automated via a network of joint locations throughout the city. Using this system people can rent a bike from one location and return it to a different place as and when needed.

Considerations: You are building a Bicycle Sharing demand forecasting service that combines historical usage patterns with weather data to forecast the Bicycle rental demand in real-time. To develop this system, you must first explore the dataset and build a model. Once it's done you must persist the model and then on each request run a Spark job to load the model and make predictions on each Spark Streaming request