PROTECHSKILLS
G14, 1st Floor, Sector 3
Near Sector 16 Metro Station
Noida, Uttar Pradesh – 201301
Website: www.protechskills.com
Tel: +91-8447-603-880

Apache Spark with Python

Overview

“Big data" analysis is a hot and highly valuable skill – and this course will teach you the hottest technology in big data: **Apache Spark**. Employers including **Amazon**, **EBay**, **NASA JPL**, and **Yahoo** all use Spark to quickly extract meaning from massive data sets across a fault-tolerant **Hadoop** cluster. You'll learn those same techniques, using your own Windows system right at home. It's easier than you might think.

Extremely Hands-On...

Incredibly Practical...

Unbelievably Real!

This course uses the familiar Python programming language.

Upon completing this course you will know:

- Learn the concepts of Spark's Resilient Distributed Datastores
- Develop and run Spark jobs quickly using Python

PROTECHSKILLS
G14, 1st Floor, Sector 3
Near Sector 16 Metro Station
Noida, Uttar Pradesh – 201301
Website: www.protechskills.com
Tel: +91-8447-603-880

- Translate complex analysis problems into iterative or multi-stage Spark scripts
- Learn about other Spark technologies, like Spark SQL, Spark Streaming, and GraphX

Targeted Audience

Students having prior knowledge of basic python and interested to choose their career as Big Data Scientist and Apache Spark developers.

Note:

The following unit and exercise durations are estimates, and might not reflect every class experience. If the course is customized or abbreviated, the duration of unchanged units will probably increase.

Course Agenda

Unit 1. Getting started with Python

Why Python?

What is Python?

Who are using Python?

Where we are using Python?

Setting up Environment.

Unit 2. Core Python

Basics of Python

Basic Data Types and Objects

Conditioning in python

Looping and breaks

Class definition on python

Unit 3. Introducing Python Modules

PROTECHSKILLS
G14, 1st Floor, Sector 3
Near Sector 16 Metro Station
Noida, Uttar Pradesh – 201301
Website: www.protechskills.com
Tel: +91-8447-603-880

Numpy

Working with Numpy.

Fast analysis and data handling with Pandas.

Exercise 1 : Average gold,silver and bronze medal problem

Pandas

Working with pandas data structures.

Working with pandas visualization.

Reading and Writing files with pandas

Matplotlib and Seaborn visualization

Working with matplotlib : creating figures and adding multiple axes.

Working with seaborn :add-on regression, distribution and matrix plots.

Activity : 3D - Plotting

Milestone Project : Titanic Survival data preprocessing

Unit 4: Introduction to Apache Spark

Why Apache Spark?

Spark Features.

Spark Ecosystem.

Environment setup.

Unit 5: Spark Basics and Simple Examples

The Resilient Distributed Dataset (RDD).

Pros and cons. Of RDDs.

Working with spark DataFrames.

Unit 6: SparkSQL

PROTECHSKILLS
G14, 1st Floor, Sector 3
Near Sector 16 Metro Station
Noida, Uttar Pradesh – 201301
Website: www.protechskills.com
Tel: +91-8447-603-880

Introduction to SparkSQL.

Executing SQL commands and SQL-style functions on a DataFrame.

Using Spark DataFrames instead of RDDs.

Unit 7: Spark MLlib

Introducing MLlib.

Using machine learning techniques in spark.

Making movie reommendations with movie lens Dataset.

Unit 8: Spark streaming

Introduction to Spark streaming.

Streaming Twitter data with Spark streaming.

Twitter top hashtags using Spark in real-time.

Ending notes : GraphX

Projects inclosed

1. Movie Recommendation using Movielens Dataset
2. Twitter Top hashtags using spark streaming in realtime

Disclaimer

All the assignments and discussion links will be provided after the lecture of current topic.