



PySpark Certification Training

Table of Contents

1. About the Program
2. About Intellipaap
3. Key Features
4. Career Support
5. Why take up this course?
6. Who should take up this course?
7. Program Curriculum
8. Certification
9. Intellipaap Success Stories
10. Contact Us



About the Program

Intellipaat's PySpark course is designed to help you understand the PySpark concept and develop custom, feature-rich applications using Python and Spark. This is one of the best PySpark training courses conducted online by leading PySpark experts working in top MNCs. During this PySpark course, you will gain in-depth knowledge of Apache Spark and related ecosystems, including Spark Framework, PySpark SQL, PySpark Streaming, and more. Also, you can work in a virtual lab and run real-time projects to get hands-on experience with PySpark.

About Intellipaat

Intellipaat is one of the leading e-learning training providers with more than 600,000 learners across 55+ countries. We are on a mission to democratize education as we believe that everyone has the right to quality education.

Our courses are delivered by subject matter experts from top MNCs, and our world-class pedagogy enables learners to quickly learn difficult topics in no time. Our 24/7 technical support and career services will help them jump-start their careers in their dream companies.

Key Features



**24 HRS INSTRUCTOR-LED
TRAINING**



22 HRS SELF-PACED TRAINING



**60 HRS REAL-TIME
PROJECT WORK**



LIFETIME ACCESS



24/7 TECHNICAL SUPPORT



**INDUSTRY-RECOGNIZED
CERTIFICATION**



**JOB ASSISTANCE THROUGH
80+ CORPORATE TIE-UPS**



FLEXIBLE SCHEDULING

Career Support



SESSIONS WITH INDUSTRY MENTORS

Attend sessions from top industry experts and get guidance on how to boost your career growth



MOCK INTERVIEWS

Mock interviews to make you prepare for cracking interviews by top employers



GUARANTEED INTERVIEWS & JOB SUPPORT

Get interviewed by our 400+ hiring partners



RESUME PREPARATION

Get assistance in creating a world-class resume from our career services team



Why take up this course?

- In the United States, Spark Developers earn an average annual salary of US\$150,000 – Neuvoo
- The average salary range for Apache Spark Developers starts from US\$92,176/year for a developer and goes up to US\$126,114/year for Back-end Developers – Indeed
- Big Data market revenue is expected to grow from US\$42 billion (2018) to US\$103 billion in 2027 – Forbes
- 79% of company executives say that companies that do not embrace Big Data are losing market control and may become non-existent – Accenture

Almost all the companies that rely on Big Data use Spark as part of their solution strategy, so job requirements in Big Data and PySpark will not reduce in the upcoming years. It is the perfect time to upskill in PySpark by enrolling in a recognized PySpark training course.

Who should take up this course?

- Software Developers, Architects, and Mainframe Developers
- ETL and DW Professionals
- BI and Senior IT Experts
- Data Science Engineers, Big Data Engineers, Developers, Architects, etc.

Program Curriculum

PySpark Training Course Content

- **INTRODUCTION TO THE BASICS OF PYTHON**
 - Explaining Python and highlighting its importance
 - Setting up the Python environment and discussing flow control
 - Running Python scripts and exploring Python editors and IDEs
- **SEQUENCE & FILE OPERATIONS**
 - Defining reserve keywords and command-line arguments
 - Describing flow control and sequencing
 - Indexing and slicing
 - Learning the xrange() function
 - Working around dictionaries and sets
 - Working with files
- **FUNCTIONS, SORTING, ERRORS & EXCEPTIONS, REGULAR EXPRESSIONS, & PACKAGES**
 - Explaining functions and various forms of function arguments
 - Learning variable scope, function parameters, and lambda functions
 - Sorting using Python
 - Exception handling
 - Package installation
 - Regular expressions
- **PYTHON: AN OOP IMPLEMENTATION**
 - Using class, objects, and attributes

- Developing applications based on OOP
- Learning about classes and objects and how they function together
- Explaining OOPs concepts, including inheritance, encapsulation, and polymorphism, among others

- **DEBUGGING & DATABASES**

- Debugging Python scripts using PDB and IDE
- Classifying errors and developing test units
- Implementing databases using SQLite
- Performing CRUD operations

- **INTRODUCTION TO BIG DATA & APACHE SPARK**

- What is Big Data?
- 5 Vs of Big Data
- Problems related to Big Data: Use Cases
- What are the tools available for handling Big Data?
- What is Hadoop?
- Why Hadoop?
- Key characteristics of Hadoop
- Important Hadoop ecosystem concepts
- MapReduce and HDFS
- Introduction to Apache Spark
- What is Apache Spark?
- Why Apache Spark?
- Who uses Spark in the industry?
- Apache Spark architecture
- Spark vs Hadoop

- Various Big Data applications using Apache Spark

- **PYTHON FOR SPARK**

- Introduction to PySpark
- Who uses PySpark?
- Why Python for Spark?
- Values, types, and variables
- Operands and expressions
- Conditional statements
- Loops
- Numbers
- Python files I/O functions
- Strings and associated operations
- Sets and associated operations
- Lists and associated operations
- Tuples and associated operations
- Dictionaries and associated operations

***Hands-on Exercise:** Demonstrating loops and conditional statements, tuples: related operations, properties, etc., lists: operations, related properties, etc., sets: properties, associated operations, etc., and dictionaries: operations, related properties, etc.*

- **PYTHON FOR SPARK: FUNCTIONAL & OBJECT-ORIENTED MODEL**

- Functions
- Lambda functions
- Global variables, their scope, and returning values
- Standard libraries

- Object-oriented concepts
- Modules used in Python
- The import statement
- Module search path
- Package installation ways

***Hands-on Exercise:** Lambda: features, options, syntax, and comparison with the functions, functions: syntax, return values, arguments, and keyword arguments, errors and exceptions: issue types and remediation, packages and modules: import options, modules, and sys path*

- **APACHE SPARK FRAMEWORK & RDDS**

- Spark components and its architecture
- Spark deployment modes
- Spark Web UI
- Introduction to PySpark Shell
- Submitting a PySpark job
- Writing your first PySpark job using Jupyter Notebook
- What are Spark RDDs?
- Stopgaps in existing computing methodologies
- How does an RDD solve the problem?
- What are the ways to create RDDs in PySpark?
- RDD persistence and caching
- General operations: Transformations, actions, and functions
- Concept of Key-Value pair in RDDs
- Other pair and two pair RDDs
- RDD lineage

- RDD persistence
- WordCount program using RDD concepts
- RDD partitioning and how it helps achieve parallelization
- Passing functions to Spark

Hands-on Exercise: Building and running a Spark application, Spark Application Web UI, loading data into RDDs, saving the data through RDDs, RDD transformations, RDD actions and functions, RDD partitions, and the WordCount program using RDDs in Python

- **PYSPARK SQL & DATAFRAMES**

- Need for Spark SQL
- What is Spark SQL?
- Spark SQL architecture
- SQL Context in Spark SQL
- User-defined functions
- DataFrames
- Interoperating with RDDs
- Loading data through different sources
- Performance tuning
- Spark–Hive integration

Hands-on Exercise: Spark SQL: Creating DataFrames, loading and transforming data through different sources, and Spark–Hive integration

- **APACHE KAFKA & FLUME**

- Why Kafka?
- What is Kafka?
- Kafka workflow
- Kafka architecture

- Kafka cluster configuring
- Kafka monitoring tools
- Basic operations
- What is Apache Flume?
- Integrating Apache Flume with Apache Kafka

Hands-on Exercise: Single-broker Kafka cluster, multi-broker Kafka cluster, topic operations, and integrating Apache Flume and Apache Kafka

- **PYSPARK STREAMING**

- Introduction to Spark Streaming
- Features of Spark Streaming
- Spark Streaming workflow
- StreamingContext initializing
- Discretized Streams (DStreams)
- Input DStreams and receivers
- Transformations on DStreams
- DStreams output operations
- Describing windowed operators and why it is useful
- Stateful operators
- Vital windowed operators
- Twitter sentiment analysis
- Streaming using the Netcat server
- WordCount program using Kafka and Spark Streaming

Hands-on Exercise: Twitter sentiment analysis, streaming using the Netcat server, WordCount program using Kafka and Spark Streaming, and Spark-Flume integration

- **INTRODUCTION TO PYSPARK MACHINE LEARNING**

- Introduction to Machine Learning: What, why, and where?
- Use cases
- Types of Machine Learning techniques
- Why use Machine Learning for Spark?
- Applications of Machine Learning (general)
- Applications of Machine Learning with Spark
- Introduction to MLlib
- Features of MLlib and MLlib tools
- Various ML algorithms supported by MLlib
- Supervised learning algorithms
- Unsupervised learning algorithms
- ML workflow utilities

Hands-on Exercise: *K-means clustering, linear regression, logistic regression, decision trees, and random forest*

Certification

After the completion of the course, you will get a certificate from Intellipaate.



CERTIFICATE OF COMPLETION

This certificate is awarded to

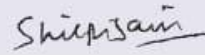
Your Name

Who has successfully completed

Course Name

Fulfilling all the requirements stipulated by Intellipaate to achieve professional excellence.

Issued Date: Month XX, XXXX



Mrs. Shilpi Jain
Director,
intellipaate Software Solutions Pvt. Ltd.

**VERIFIED
CERTIFICATE**

Certificate ID #94658291

Success Stories



Kevin K Wada

Thank you very much for your top-class service. A special mention should be made for your patience in listening to my queries and giving me a solution, which was exactly what I was looking for. I am giving you a 10 on 10!



Sampson Basoah

The Intellipaate team helped me in selecting the perfect course that suits my profile. The whole course was practically oriented, and the trainers were always ready to answer any question. I found this course to be impactful. Thank you.



Sugandha Sinha

Intellipaate's course instructors were excellent and well-versed with their concepts. The support team solved all my queries within the promised 24 hours. They explained all topics and concepts well, and the course material was updated and included videos, exercises, etc. I would highly recommend Intellipaate to those who wish to excel in the IT field.



Vishal Pentakota

The best part of this course was the series of hands-on demonstrations that the trainer performed. Not only did he explain each concept theoretically, but he also implemented all those concepts practically. Great job! A must go for beginners.

CONTACT US

INTELLIPAAT SOFTWARE SOLUTIONS PVT. LTD.

Bangalore

AMR Tech Park 3, Ground Floor, Tower B,
Hongasandra Village, Bommanahalli,
Hosur Road, Bangalore – 560068

USA

1219 E. Hillsdale Blvd. Suite 205,
Foster City, CA 94404

If you have any further queries or just want to have a conversation with us, then do call us.

IND: +91-7022374614 | US: 1-800-216-8930