PySpark

Process and analyse large datasets with PySpark, from RDDs and DataFrames to streaming.

DURATION	FORMAT	LEVEL	LABS
45 hours	Onsite, online or hybrid	Intermediate	Cloud sandbox included

About this programme

A hands-on programme on Apache Spark's Python API: Spark architecture and its Hadoop foundations, a Python refresher, RDD and DataFrame programming, cluster execution and tuning, and streaming with Kafka.

Who it is for: Python developers, data engineers and analysts working with large datasets

What your team will be able to do

- Explain Spark's architecture, job execution flow and cluster managers
- Write PySpark code with RDDs, DataFrames, joins and aggregations
- Use caching, broadcast variables and accumulators effectively
- Run ETL jobs against file systems and databases with spark-submit
- Build streaming pipelines with Spark and Kafka

Before you start

- Working Python
- Working SQL

Course outline

1. Big data and Spark foundations

- Big data analytics and Spark use cases in industry
- Spark architecture, job execution flow and why Spark with Python
- HDFS and MapReduce: how distributed processing works
- The data analytics lifecycle: discovery to insights

2. Python refresher

- Control flow, strings, lists, tuples, sets and dictionaries
- Functions, lambdas, modules and regular expressions
- File I/O, exceptions, date and time, and multithreading

3. PySpark programming

- SparkContext and environment set-up
- RDDs, caching, transformations and actions
- Key-value pairs, aggregate functions and joins
- Spark DataFrames

4. Advanced Spark programming

- Shared variables: broadcast variables and accumulators
- Custom accumulators and fault tolerance
- Numeric RDD and per-partition operations

5. Running Spark on a cluster

- Runtime architecture: driver, executors and cluster managers
- ETL from file systems and databases
- Storage levels and serialisers
- spark-submit and performance tuning

6. Spark Streaming

- Introduction to Spark Streaming
- PySpark streaming with Apache Kafka
- Operations and window functions on streaming DataFrames
- Real-world streaming use cases

Sample schedule

6 days onsite, or 12 half-day sessions online. We fit days and session lengths to your team's calendar.

Day 1	8 h	Big data and Spark foundations
Day 2	8 h	Python refresher; PySpark programming
Day 3	8 h	PySpark programming (continued); Advanced Spark programming
Day 4	8 h	Advanced Spark programming (continued); Running Spark on a cluster
Day 5	8 h	Running Spark on a cluster (continued); Spark Streaming
Day 6	5 h	Spark Streaming (continued)

Assessment and certification

How learning is checked

- A short knowledge check on day one and again at the end, so you can see each person's progress
- Graded hands-on labs in every module, reviewed by the trainer
- A final exercise that applies the programme to a problem like your team's own

Certificate and exams

- Participants who complete the programme receive a Cognitive Networks AI certificate of completion.

Formats and cohort size

- Onsite: At your office or a venue you choose, in full days.
- Online, live: Instructor-led over video, usually in half-day sessions so people can keep up with work.
- Hybrid: Some days in a room together, the rest online. Useful when a team works across cities.

Cohort size: 15 to 25 people per trainer. Larger teams run as parallel cohorts on the same outline, so everyone learns the same thing.

Pricing: We quote per cohort, based on duration, format and number of cohorts. There is no per-seat price list: tell us your team and we send a proposal within one business day.

Bring this programme to your team

Request a proposal at cognitivenetworks.ai/contact?programme=pyspark

reachus@cognitivenetworks.ai · Bangalore +91 72043 05303 · Dallas +1 (303) 827-7415