SYLLABUS FOR BOARD OF STUDIES ADOPTION — generated from the live catalogue record

Big Data Processing with Apache Spark

S-12A · 5 days (configurable 1–5 days) · Informatics · Data & Analytics · Years: 3rd, PG

CONTENTS

Catalogue S-12A · version 1 · generated 2026-08-24

OVERVIEW

Distributed processing, DataFrames, Spark SQL and a pipeline over a genuinely large dataset.

When data outgrows a single machine, Spark is the industry answer; hands-on distributed processing over a genuinely large dataset is a scarce, valuable skill.

HOURS AND CREDIT POSITION

Contact hours: 40 Guided project: 20 Self-study: 0 Notional total: 60

Credit claim: 2
60 notional hours ÷ 30 hours per credit (NSQF track, UGC Micro-credentials Guidelines §5.3.3) = 2 credits.

COURSE OUTCOMES

CO1. Explain the distributed-processing model that Spark implements (K2)
CO2. Transform data using the Spark DataFrame API (K3)
CO3. Query large datasets with Spark SQL (K3)
CO4. Optimise a Spark job through partitioning and caching (K4)
CO5. Build a Spark pipeline over a large dataset (K3)

COURSE CONTENT — MODULES

Hours shown are the recommended format; module time scales proportionally when the engagement runs 1–5 days.

Module 1 — Big data foundations (5h recommended)
Theory: The distributed-processing problem, the Spark model
Lab: Reason about why single-machine processing fails
Output: Foundations note

Module 2 — Spark core (5h recommended)
Theory: RDDs, transformations, actions, lazy evaluation
Lab: Transform data with RDDs
Output: Core notebook

Module 3 — DataFrames (5h recommended)
Theory: Structured data, the DataFrame API
Lab: Process structured data with DataFrames
Output: DataFrame notebook

Module 4 — Spark SQL (5h recommended)
Theory: Querying at scale, query optimisation
Lab: Query a large dataset with Spark SQL
Output: Query notebook

Module 5 — Data sources and formats (5h recommended)
Theory: Partitioning, columnar formats
Lab: Read and write partitioned formats
Output: Format exercise

Module 6 — Performance (5h recommended)
Theory: Partitioning, caching, shuffles
Lab: Optimise a slow Spark job
Output: Optimised job

Module 7 — Streaming basics (5h recommended)
Theory: Structured streaming
Lab: Build a simple streaming job
Output: Streaming job

Module 8 — Capstone (5h recommended)
Theory: A pipeline over a genuinely large dataset
Lab: Build the capstone pipeline
Output: Large-scale pipeline

ASSESSMENT SCHEME (100 MARKS)

Continuous lab exercises — 40 marks
Capstone project & demonstration — 40 marks
Quiz & viva voce — 20 marks

CO–PO MAPPING

CO1 !' PO1 (strength 2)
The distributed model is foundational knowledge.

CO2 !' PO5 (strength 2)
The DataFrame API is a modern tool.

CO3 !' PO5 (strength 2)
Spark SQL is a modern tool.

CO4 !' PO4 (strength 2)
Performance tuning is investigation of job behaviour.

CO5 !' PO3 (strength 2)
The pipeline is a developed solution.

EVERY PARTICIPANT RECEIVES

- Certificate of completion
- Course material
- LMS access
- Interview question bank
- Mock interview & viva practice
- Optional nasscom NSQF assessment (priced per candidate on your quotation)

FEE

[1]3,200 per student, incl. GST
Excludes the optional nasscom assessment where offered. Final pricing on your college's quotation.

DESIGNED BY

Venkat Kurela — Founder & Director, Kurela Cognisive Private Limited