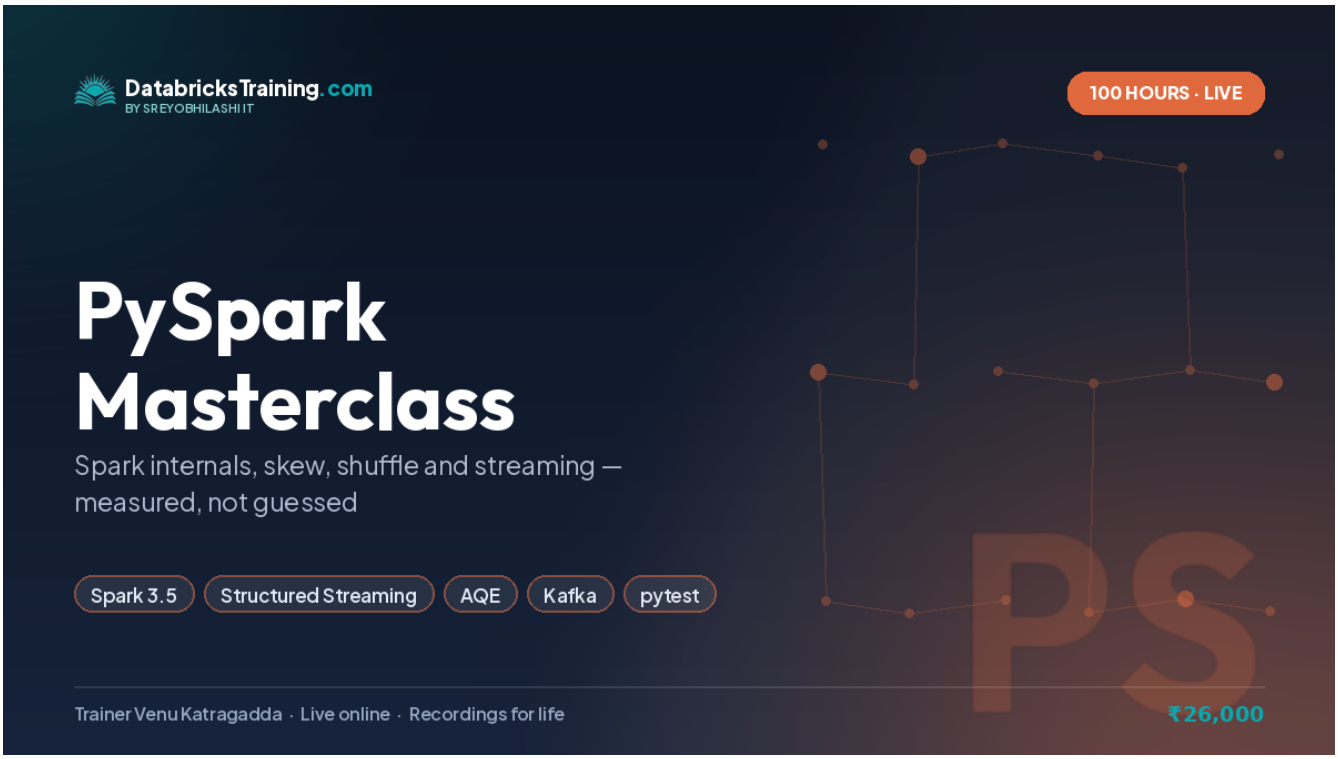

COURSE SYLLABUS

PySpark Masterclass

The deepest PySpark course we teach — internals, tuning, testing and streaming, not just the DataFrame API.

100 hours	10 modules	3 projects	**₹26,000**
LIVE INSTRUCTION	STRUCTURED PATH	PORTFOLIO READY	EMI AVAILABLE

Level	Beginner to Advanced
Mode	Live, instructor-led online sessions. Every session recorded; lifetime access.
Trainer	Venu Katragadda — 14+ years in industry (L&T Infotech, Capgemini, Brillio, Manthan Systems). 1,200+ professionals trained.
Batch timings	Morning 6:00–7:30 AM IST · Morning 7:30–9:00 AM IST · Evening 7:30–9:00 PM IST · Weekend (Sat & Sun)
Certification	Databricks Certified Associate Developer for Apache Spark
Enquiries	+91-8500002025 / +91-9247159150 (WhatsApp) · info@databrickstraining.com

What you will be able to do

Every outcome below is demonstrated in a lab or a graded project — not just discussed in a slide.

- Write idiomatic PySpark that the Catalyst optimiser can actually optimise
- Read a Spark UI and fix skew, spill, small files and shuffle problems on your own
- Build streaming pipelines with exactly-once guarantees and stateful operators
- Unit-test Spark code with pytest and chispa and ship it through CI/CD
- Run the same codebase on Databricks, AWS EMR, GCP Dataproc and Kubernetes
- Answer the Spark internals questions that decide senior data engineering interviews

Who this masterclass is for

- Python developers moving into big data
- Data engineers stuck at the 'it works but it's slow' stage
- Data scientists who need production-grade pipelines
- Anyone preparing for Spark developer interviews

Prerequisites

- Comfortable with Python basics (or take the 8-hour Python primer in Module 1)
- Basic SQL
- 8 GB RAM laptop; Docker used for local Spark

Tools and technologies

Apache Spark 3.5/4.0 · PySpark · Spark SQL · Structured Streaming · Delta Lake · Kafka · Airflow · pytest · chispa · Docker · AWS EMR · GCP Dataproc · Databricks

Curriculum — 100 hours across 10 modules

Sessions run live in small batches. Roughly 60% of each session is hands-on lab work.

MODULE 1 Python primer for data engineers — 8 hrs

- Data types, comprehensions, generators, decorators, typing
- Modules, packaging, virtual environments, uv/poetry, logging
- Working with JSON, CSV, Parquet in pandas; pandas ↔ Spark interop
- Lab: build a small ETL utility package

MODULE 2 Spark architecture and execution model — 10 hrs

- Cluster managers: standalone, YARN, Kubernetes, Databricks
- Driver/executor memory model, cores, slots, dynamic allocation
- RDD lineage, DAG, jobs/stages/tasks, narrow vs wide transformations
- Catalyst optimiser, Tungsten, whole-stage codegen, AQE
- Lab: run the same job on 1, 4 and 16 cores and explain the curve

MODULE 3 DataFrames and Spark SQL in depth — 14 hrs

- Schemas: StructType, nested structs, arrays, maps, explode/flatten
- All join types and strategies: broadcast hash, sort-merge, shuffle hash, bucketed
- Window functions, ranking, running totals, sessionisation
- UDFs vs pandas UDFs vs Spark built-ins — measured performance comparison
- Handling nulls, dedup, data-quality checks, quarantine patterns
- Lab: flatten a deeply nested JSON API payload into a star schema

MODULE 4 File formats, storage and partitioning — 8 hrs

- Parquet internals: row groups, column chunks, statistics, predicate pushdown
- Avro, ORC, JSON, Delta, Iceberg — when each wins
- Partitioning vs bucketing; the small-file problem and how to fix it
- Compression codecs: snappy vs zstd vs gzip benchmarks
- Lab: cut a table's scan time 6x by re-laying-out the files

MODULE 5 Structured Streaming — 12 hrs

- Micro-batch vs continuous; sources and sinks; foreachBatch
- Checkpointing, offsets, idempotent writes, exactly-once semantics
- Event time, watermarks, windowing (tumbling, sliding, session)
- Stream-stream and stream-static joins; state store internals and RocksDB
- Kafka integration: consumer groups, maxOffsetsPerTrigger, back-pressure
- Lab: streaming sessionisation with late-arriving events

MODULE 6 Performance tuning — the deep session — 14 hrs

- Reading the Spark UI end to end: stages, tasks, SQL tab, executor tab
- Diagnosing skew and fixing it: salting, AQE, iterative broadcast
- Spill, GC pressure, memory fractions, off-heap, serialization (Kryo)
- Shuffle partitions: the maths behind spark.sql.shuffle.partitions

- Caching and persistence levels — when caching makes things worse
- Cost-based optimisation, statistics, ANALYZE TABLE
- Lab: three broken jobs, three root-cause reports, three fixes

MODULE 7 Testing, packaging and code quality — 8 hrs

- pytest fixtures for SparkSession; chispa DataFrame assertions
- Property-based testing, fixtures from JSON, golden datasets
- Project layout, config with pydantic/YAML, dependency injection
- Linting and typing: ruff, mypy; pre-commit hooks
- Lab: retrofit tests onto an untested legacy pipeline

MODULE 8 Orchestration and deployment — 10 hrs

- spark-submit anatomy: deploy modes, --packages, --conf, --files
- Running on AWS EMR (EMR Serverless too), GCP Dataproc, Kubernetes (Spark Operator)
- Airflow DAGs for Spark: SparkSubmitOperator, EmrOperators, sensors, retries
- CI/CD: build, test, publish artefact, deploy, smoke test
- Lab: promote a job through dev → staging → prod with GitHub Actions

MODULE 9 Advanced and adjacent topics — 8 hrs

- Spark Connect and Spark 4.0 changes; PySpark on pandas API
- Arrow, vectorisation, Photon and native engines (Comet, Gluten, Velox)
- Graph processing with GraphFrames; ML pipelines with MLlib overview
- Iceberg and Hudi with Spark: table formats compared
- Common anti-patterns: collect(), toPandas(), row-by-row loops

MODULE 10 Capstone, interview prep and certification — 8 hrs

- Capstone build and architecture review
- 80+ Spark interview questions with model answers, from junior to architect level
- Live coding drill: 12 timed PySpark problems
- Databricks Certified Associate Developer for Apache Spark exam walkthrough
- Resume writing for Spark roles and how to describe your projects

Hands-on projects

You leave with work you can demo in an interview. Each project is code-reviewed and returned with written feedback.

PROJECT 1 NYC taxi 1.5 B rows

Full ETL over the public taxi dataset: partitioning strategy, broadcast joins, skew handling, and a benchmark report showing a 9x speedup.

PROJECT 2 Real-time fraud scoring

Kafka → Structured Streaming → stateful rules engine → Delta sink, with watermarking, late data handling and replay.

PROJECT 3 Tested and packaged PySpark app

A pip-installable PySpark project with config management, pytest suite, GitHub Actions CI and spark-submit deployment to EMR.

Frequently asked questions

How is this different from the Databricks course?

This course is about Spark itself — internals, tuning, streaming and testing — and runs on Databricks, EMR, Dataproc and local Docker. The Databricks course is about the platform: Unity Catalog, DLT, Workflows, SQL warehouses and governance. Many students take both; there is a bundle discount.

Which Spark version do you teach?

Spark 3.5 as the baseline with a dedicated session on Spark 4.0 changes (Spark Connect, ANSI mode by default, VARIANT type). Deprecated APIs are called out explicitly as we go.

Do I need a cloud account?

Not for the first half — labs run in Docker locally. From Module 8 you'll use free-tier AWS/GCP credits or Databricks Community Edition; we walk through setup and cost guardrails.

Is Scala covered?

Only for reading. The course is Python-first, but you'll learn to read Scala Spark code because much of the ecosystem's source and Stack Overflow answers are in Scala.

How to enrol

- Book a free demo. Sit in on a live session with the current batch — no card, no commitment. WhatsApp +91-9247159150 or call +91-8500002025.
- Pick your batch. Weekday morning, weekday evening or weekend. All times IST.
- Pay the fee (₹26,000, instalments and EMI available) and get your workspace, material and calendar invites.
- Start building. Labs from session one; recordings posted the same day.

What is included

- 100 hours of live instruction by Venu Katragadda
- Lifetime access to all session recordings
- 3 code-reviewed portfolio projects
- Preparation for: Databricks Certified Associate Developer for Apache Spark
- Resume review, LinkedIn optimisation and one mock interview with written feedback
- Free re-attend of any future batch of the same masterclass
- Doubt-clearing sessions and a private student group

Refund policy in brief

When you cancel	Refund
Before the batch starts	Up to 90%
Within 5 days of start, max 2 classes attended	Up to 70%
After 7 days, or more than 5 classes attended	No refund — free transfer to a later batch

Full policy: databrickstraining.com/refund-policy/

Talk to us

Call	+91-8500002025 · +91-9247159150
WhatsApp	+91-9247159150 — usually answered within the hour
Email	info@databrickstraining.com
Website	databrickstraining.com/courses/pyspark-masterclass/
Foundation tier	databrickstraining.in — shorter 70–80 hour courses if this one is too advanced
Institute	Sreyobhilashi IT · Udyam UDYAM-TS-02-0334643 · 112, Samshraya Homes, opp. PF Office, Balanagar, Kukatpally, Hyderabad, Telangana 500037