

Apache Spark & PySpark

Course Duration ~75 Hours · 10 Weeks · Live Online · Lab-Heavy · Portfolio of Real Pipelines

Spark 4.x

DataFrames & SQL

Delta Lake

Streaming

Databricks

Performance

In short: Process terabytes without melting the cluster — or the budget. This is the data engineering foundation of our catalogue: the pipelines that feed analytics, ML and every AI system downstream — and the one course whose core technology has been stable for a decade. It targets the current release line, not 2017 tutorials: Spark 4.x with DataFrames/Spark SQL as the primary interface (RDDs deprecated, taught as background only), Adaptive Query Execution on by default, Spark Connect for local development against remote clusters, Delta Lake / Iceberg / Hudi as the normal storage layer, and Databricks — the name most job ads now use. Across 10 modules you ship medallion batch pipelines, incremental MERGE upserts, Kafka-to-Delta streaming with exactly-once guarantees, a documented performance rescue, and an orchestrated, cloud-deployed production job — closing with a batch + streaming lakehouse capstone. Built for fresh IT graduates (Python + SQL is the shortest on-ramp into data engineering) and for working professionals modernising onto the lakehouse. Aligned to the Databricks Certified Associate Developer for Apache Spark (Python).

Duration	10 weeks · ~75 hours · live + hands-on	Format	Live online · lab-heavy · portfolio of real pipelines
Prerequisite	Working Python and comfortable SQL. No Hadoop, Scala or distributed-systems background required.	Curriculum	10 modules, from distributed foundations to MLlib and capstone
Certification	Aligned to the Databricks Certified Associate Developer for Apache Spark (Python); substantial ground for the Databricks Certified Data Engineer Associate	Ideal for	Fresh IT graduates and developers/analysts/ETL professionals moving into lakehouse data engineering

When NOT to Use Spark

For a few GB a day, a warehouse or a single-node engine is the right answer — reaching for a cluster reflexively is how budgets melt. Knowing that boundary is what separates an engineer from a tool operator, and reasoning about it out loud is a common senior-level interview question. This course teaches the boundary explicitly: you will learn to justify *not* using Spark as confidently as using it.

Course Highlights

- Every module leads with what you build and ship, not what you "learn about"
- Spark 4.x throughout: DataFrames first, RDDs as background only, AQE on by default
- Delta Lake, Iceberg and Hudi as the normal storage layer — the lakehouse, not HDFS-era Hadoop
- Structured Streaming from Kafka to Delta with watermarking and exactly-once guarantees
- A dedicated performance-rescue lab: cut a skewed job's runtime with before/after plans
- Databricks in depth — the name most job ads use — plus EMR, Dataproc and Azure
- Production engineering: testing, CI/CD, orchestration, idempotent restartable jobs

Hands-on Projects & Portfolio

- **Medallion Batch Pipeline** — multi-hundred-GB raw ingest → cleaned → aggregated on Delta Lake, bronze/silver/gold
- **Incremental Orders Pipeline** — MERGE upserts, late-arriving data handling and a data quality gate
- **Kafka → Delta Streaming** — real-time pipeline with watermarking and exactly-once guarantees
- **Performance Rescue** — take a deliberately slow, skewed job and cut its runtime, with before/after execution plans and cost
- **Orchestrated Production Job** — Airflow orchestration with tests, CI/CD and failure alerting
- **Cloud-Deployed Pipeline** — EMR / Dataproc / Databricks with autoscaling and cost analysis

Tools & Stack

Python · PySpark · Spark SQL · Apache Spark 4.x · Spark Connect · Delta Lake · Apache Iceberg · Apache Kafka · Databricks · Unity Catalog · Apache Airflow · Parquet · AWS EMR · S3 · GCP Dataproc · Azure Databricks · Docker · Git

Certification Alignment

- Syllabus mapped to the **Databricks Certified Associate Developer for Apache Spark (Python)**
- Covers substantial ground for the **Databricks Certified Data Engineer Associate**
- Presented as preparation and alignment — no official partnership or accreditation implied, and no exam outcome guaranteed

Career Outcomes

Spark remains the default engine wherever there's real data volume, messy pipelines, lakehouse tooling or cloud platforms — and Python plus SQL is the shortest on-ramp into it. Many job ads now say "Databricks" rather than "Spark," which is why this course covers both. It targets:

- Data Engineer · Big Data Engineer · Analytics Engineer · Databricks Engineer
- ETL Developer · Data Platform Engineer · Machine Learning Engineer (data-heavy) · Senior Data Analyst working beyond warehouse scale

Who This Course Is For

- Final-year students and fresh IT graduates targeting data engineering roles
- Python developers and software engineers moving into data engineering
- SQL-strong data analysts and BI professionals hitting the limits of warehouse-only tooling
- ETL and ODI/Informatica-style developers modernising onto the lakehouse
- Data scientists who need their work to run on production-scale data
- Engineers preparing for Databricks certification

Module 1: Distributed Data Foundations

7 Hours

YOU SHIP

- Your first distributed job, traced end-to-end through the Spark UI — jobs, stages, tasks and shuffles annotated by hand

TOPICS

- Why single-machine processing breaks; the Spark execution model; driver and executors
- Jobs/stages/tasks; lazy evaluation; partitions and shuffles
- Cluster managers (YARN, Kubernetes, Databricks); reading the Spark UI from day one

Assessment: Quiz + Lab

Module 2: PySpark DataFrames & Spark SQL

9 Hours

YOU SHIP

- A multi-source transformation pipeline — joins, window functions, nested types — written in the API you'll use for 80% of real work, with explicit schemas throughout

TOPICS

- Reading and writing every common format; explicit schemas instead of inference
- Transformations and actions; joins; window functions; aggregations; complex and nested types
- Mixing SQL with DataFrame code cleanly

Assessment: Quiz + Lab

Module 3: Working with Real, Messy Data

7 Hours

YOU SHIP

- An incremental load pipeline with upserts, deduplication and late-arriving data handling — guarded by a data quality layer that fails loudly instead of silently

TOPICS

- Semi-structured JSON and XML; nested flattening; null and type handling
- Deduplication; slowly changing dimensions; incremental loads and upserts; late-arriving data
- A data quality layer that fails loudly

Assessment: Quiz + Lab**Module 4: Performance Tuning**

9 Hours

YOU SHIP

- A performance rescue: a deliberately slow, skewed job cut down to size, documented with before/after execution plans and cost

TOPICS

- Catalyst optimizer and Tungsten; reading execution plans; Adaptive Query Execution
- Partition sizing; repartition vs coalesce; broadcast joins; caching and persistence levels
- Diagnosing and fixing data skew; avoiding Python UDFs in favour of built-ins; Pandas UDFs with Arrow when you truly need them

Assessment: Quiz + Lab**Module 5: Delta Lake & Open Table Formats**

8 Hours

YOU SHIP

- A transactional Delta Lake layer with MERGE upserts, schema evolution, time travel rollback and an OPTIMIZE/compaction routine

TOPICS

- ACID transactions on object storage; schema enforcement and evolution; MERGE for upserts
- Time travel and rollback; OPTIMIZE, Z-ordering and compaction; vacuum
- How Apache Iceberg and Hudi compare and when each is chosen

Assessment: Quiz + Lab**Module 6: Structured Streaming**

8 Hours

YOU SHIP

- A real-time pipeline from Kafka into Delta with watermarking, stateful aggregations, exactly-once semantics and checkpointed recovery

TOPICS

- The same DataFrame API applied to real-time data; Kafka sources and sinks; micro-batch vs continuous
- Event time vs processing time; watermarking; stateful aggregations
- Exactly-once semantics; checkpointing and recovery; streaming into Delta

Assessment: Quiz + Lab**Module 7: Databricks & the Lakehouse**

7 Hours

YOU SHIP

- A governed medallion (bronze/silver/gold) pipeline on Databricks — declarative pipelines, Unity Catalog lineage, scheduled workflows

TOPICS

- Workspace, clusters and notebooks; Databricks SQL
- Delta Live Tables / declarative pipelines; Unity Catalog for governance and lineage
- Jobs and workflows; medallion architecture in practice

Assessment: Quiz + Lab

Module 8: Spark on the Cloud

6 Hours

YOU SHIP

- The same pipeline deployed to a managed cloud cluster — packaged, submitted, autoscaled on spot instances, with a cost analysis attached

TOPICS

- AWS EMR, GCP Dataproc and Azure; object storage access patterns; spark-submit and packaging
- Spark Connect for local development against remote clusters
- Cluster sizing; autoscaling; spot instances and cost control

Assessment: Quiz + Lab**Module 9: Production Engineering**

7 Hours

YOU SHIP

- A production-grade Spark job — tested, logged, Airflow-orchestrated, idempotent and restartable — with CI/CD and failure alerting

TOPICS

- Structuring a Spark repo; testing PySpark code; logging and the Spark 4 structured logging framework
- Orchestration with Airflow; CI/CD for pipelines; idempotent and restartable jobs
- Monitoring, alerting and on-call for data pipelines

Assessment: Quiz + Lab**Module 10: Spark MLlib & Capstone**

7 Hours

YOU SHIP

- One end-to-end batch + streaming lakehouse pipeline — designed, built and defended — plus a distributed ML pipeline with cross-validation at scale

TOPICS

- Distributed feature engineering; ML pipelines and cross-validation at scale
- When to use MLlib versus handing off to a Python ML stack
- Capstone design, build and defense

Assessment: Capstone Project + Review