

Databricks Masterclass

Build a production Lakehouse and tune what it costs to run

Delta Lake Unity Catalog DLT Photon Workflows

Trainer Venu Katragadda · Live online · Recordings for life

₹32,000

COURSE SYLLABUS

Databricks Masterclass

Go from SQL/Python basics to a production Lakehouse you built yourself — Delta Lake, Unity Catalog, DLT, Workflows and cost governance.

100 hours LIVE INSTRUCTION	10 modules STRUCTURED PATH	3 projects PORTFOLIO READY	₹32,000 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 Data Engineer Associate + Professional
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.

- Design a medallion (bronze/silver/gold) Lakehouse on Delta Lake that survives schema drift and late-arriving data
- Build incremental pipelines with Auto Loader, Structured Streaming and Delta Live Tables
- Govern data with Unity Catalog: catalogs, external locations, row/column masking and lineage
- Tune jobs with the Spark UI, Photon, liquid clustering, Z-ORDER and predictive optimisation
- Cut cluster spend with job compute, spot fleets, autoscaling policies and system-table chargeback
- Pass the Databricks Certified Data Engineer Associate and Professional exams

Who this masterclass is for

- Data engineers and ETL/Informatica developers moving to the Lakehouse
- SQL developers, DBAs and BI engineers modernising a warehouse
- Python/Java developers entering big data
- Architects preparing a Databricks migration

Prerequisites

- Basic SQL (SELECT, JOIN, GROUP BY)
- Any one programming language — Python is taught from scratch in Module 1
- A laptop with 8 GB RAM; free Databricks Community Edition + trial cloud accounts are used in labs

Tools and technologies

Databricks · Delta Lake · Unity Catalog · Apache Spark · PySpark · Delta Live Tables · Databricks SQL · Databricks Asset Bundles · MLflow · Apache Airflow · Kafka · Terraform · GitHub Actions

Curriculum — 100 hours across 10 modules

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

MODULE 1 Foundations: Python, SQL and the Databricks workspace

8 hrs

- Python for data engineering: collections, comprehensions, functions, error handling, virtualenv/uv
- Advanced SQL refresher: window functions, CTEs, grouping sets, set operations
- Workspace tour: notebooks, Repos/Git folders, DBFS vs volumes, secrets, workspace files
- Cluster types: all-purpose vs job vs SQL warehouse vs serverless; runtime (DBR) versions and Photon
- Lab: first cluster, first notebook, connect Databricks to a GitHub repo

MODULE 2 Apache Spark architecture and the DataFrame API

12 hrs

- Driver, executors, jobs → stages → tasks; lazy evaluation and the Catalyst optimiser
- Adaptive Query Execution, dynamic partition pruning, whole-stage codegen, Photon vectorisation
- DataFrame API: select, filter, withColumn, joins, aggregations, UDFs vs pandas UDFs vs built-ins
- Spark SQL, temp views, catalogs; reading JSON, CSV, Parquet, Avro, XML, fixed-width
- Reading the Spark UI: skew, spill, shuffle read/write, GC time
- Lab: rewrite a slow 12-minute join to under 90 seconds

MODULE 3 Delta Lake deep dive

12 hrs

- Why Delta: ACID on object storage, the `_delta_log`, checkpoints and protocol versions
- MERGE, UPDATE, DELETE; SCD Type 1 and Type 2 patterns; deletion vectors
- Time travel, RESTORE, CLONE (shallow/deep), CDF (Change Data Feed)
- OPTIMIZE, Z-ORDER, liquid clustering, VACUUM, predictive optimisation, file sizing
- Schema enforcement vs schema evolution; constraints and generated columns
- Delta Sharing and open-format interop (Iceberg/UniForm)
- Lab: build a slowly-changing customer dimension with MERGE + CDF

MODULE 4 Ingestion: Auto Loader, batch and streaming

10 hrs

- Auto Loader (cloudFiles): file notification vs directory listing, schema inference and hints, rescued data
- Structured Streaming: triggers (availableNow, processingTime), checkpoints, watermarks, output modes
- Stateful streaming: joins, dropDuplicatesWithinWatermark, flatMapGroupsWithState
- Kafka + Databricks: offsets, exactly-once semantics, Avro/Protobuf with Schema Registry
- CDC ingestion with Debezium and APPLY CHANGES INTO
- Lab: exactly-once Kafka → Delta pipeline with replay and back-pressure handling

MODULE 5 Delta Live Tables and declarative pipelines

8 hrs

- DLT concepts: streaming tables, materialised views, expectations, pipeline modes
- Development vs production, continuous vs triggered, Enhanced Autoscaling
- Data-quality expectations: expect, expect_or_drop, expect_or_fail; quarantine tables
- APPLY CHANGES INTO for CDC and SCD2 without hand-written MERGE
- Lakeflow Declarative Pipelines: what changed and how to migrate
- Lab: convert a hand-rolled notebook pipeline to DLT and add quality gates

MODULE 6 Unity Catalog, security and governance

8 hrs

- Metastore, catalog, schema, table, volume, model; three-level namespace
- Managed vs external tables, external locations and storage credentials
- Grants, dynamic views, row filters and column masks for PII
- Lineage, audit logs, system tables, tags and Data Classification
- Delta Sharing to partners; catalog binding and workspace isolation
- Lab: implement GDPR-safe access for a customer table across three roles

MODULE 7 Databricks SQL and BI serving

8 hrs

- SQL warehouses: classic vs pro vs serverless, scaling, query queuing, result cache
- Query profile analysis, materialised views, caching strategies
- Dashboards, alerts, parameters; Power BI / Tableau / Looker connectivity
- Modelling the gold layer: star schema, surrogate keys, aggregate tables
- Lab: sub-second executive dashboard on a 500M-row fact table

MODULE 8 Orchestration, DevOps and CI/CD

10 hrs

- Databricks Workflows: tasks, dependencies, job clusters, retries, repair-run, job parameters
- Apache Airflow with the Databricks provider; when to use Airflow vs Workflows
- Databricks Asset Bundles (DABs): bundle.yml, targets, dev/stage/prod promotion
- CI/CD with GitHub Actions / Azure DevOps: unit tests with pytest + chispa, integration tests
- Terraform provider for workspaces, clusters, policies and permissions
- Lab: one-command deploy of a pipeline from dev to prod via a bundle

MODULE 9 Performance tuning and cost optimisation

10 hrs

- Diagnosing skew: salting, AQE skew join, broadcast thresholds
- Partitioning vs liquid clustering vs Z-ORDER — what to use when
- Caching, disk cache, Photon eligibility, serverless compute trade-offs
- Cluster policies, spot/pre-emptible fleets, autotermination, instance pools
- System tables for DBU chargeback; tagging strategy and budget alerts
- Lab: reduce a nightly job from 55 minutes / \$180 to 14 minutes / \$38

MODULE 10 Lakehouse architecture, AI and capstone

14 hrs

- Medallion architecture in practice; data mesh and domain ownership on Unity Catalog
- Migration playbooks: Hadoop/Hive, Teradata, Informatica and Synapse → Databricks
- MLflow 3, Feature Store, Model Serving and Vector Search — what a data engineer must know
- Databricks Apps, Genie and AI/BI: serving analytics to business users
- Capstone build, code review, architecture defence and demo day
- Certification drill: 120 practice questions for Associate + Professional

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 Retail Lakehouse end-to-end

Ingest 40M+ POS rows from S3/ADLS with Auto Loader, build bronze→silver→gold with DLT expectations, serve a Databricks SQL dashboard and alert on data quality drops.

PROJECT 2 Streaming clickstream + CDC

Kafka clickstream joined with Debezium CDC from Postgres, exactly-once Structured Streaming into Delta, SCD Type 2 with MERGE, watermarks and stateful aggregations.

PROJECT 3 Cost & governance audit

Use system tables to build a chargeback dashboard, apply cluster policies, Unity Catalog masking for PII, and cut a workload's DBU cost by 40%.

Frequently asked questions

Do I need Spark experience before joining?

No. Modules 1–2 build Python and Spark fundamentals from scratch. If you already know Spark you can skip ahead — recordings are released from day one.

Which cloud do the labs run on?

You can follow along on AWS, Azure or GCP. Labs are written to be cloud-neutral, with cloud-specific notes for storage and identity. Databricks Community Edition plus free cloud tiers cover most exercises.

Does this cover the Databricks certifications?

Yes. The course maps to the Databricks Certified Data Engineer Associate and Professional exam guides, and includes 120 practice questions plus two timed mock exams. The exam fee itself is paid directly to Databricks.

What if I miss a live session?

Every session is recorded and posted within a few hours. You also get lifetime access to recordings and can rejoin any future batch at no cost.

Is there job support?

Yes — resume review, LinkedIn optimisation, a mock interview and an interview question bank are included. We do not guarantee placement, and we will never sit in an interview on your behalf.

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 (₹32,000, 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 Data Engineer Associate + Professional
- 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/databricks-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

Sreyobhilashi IT is an independent training provider and is not affiliated with, endorsed by or sponsored by Databricks, Amazon, Microsoft, Google, Snowflake, dbt Labs or Anthropic. Placement assistance is a best-effort service, not a guarantee of employment. Cloud platforms change frequently; confirm current pricing and service behaviour with the vendor's documentation.