Cloud Contraptions
TRAINING & CONSULTING

To discuss this course and customizations:
Call: +1 434-509-6890 or Email: sales@cloudcontraptions.com

Advanced Machine Learning and Data Engineering

Class Duration

35 hours of live training delivered over 5 days.

Student Prerequisites

- Completion of AI and Modern Machine Learning for Software Developers or equivalent experience with ML fundamentals (preparing data, then training, evaluating, and tuning models in Python)
- Solid Python programming skills, including pandas and scikit-learn
- Solid SQL skills: joins, aggregations, and analytic queries

Target Audience

Machine learning engineers, data engineers, and software developers who need to take machine learning from working prototypes to production scale. This course builds directly on AI and Modern Machine Learning for Software Developers, extending its modeling skills with the data platform and ML operations engineering that production systems demand. Teams that only need the pipeline side without the ML engineering can instead start with our Data Engineering courses, such as Practical Apache Spark for Data Pipelines.

Description

This hands-on course is for teams who have mastered machine learning fundamentals and need to make ML work at production scale. The data platform comes first: participants build reliable pipelines with Apache Spark 4, SQL-based transformation with dbt, orchestration with Apache Airflow 3, streaming ingestion with Apache Kafka 4, and lakehouse architecture on the open table formats Apache Iceberg and Delta Lake. On that foundation sits the advanced ML engineering: feature pipelines with point-in-time correctness, distributed training, experiment tracking and a model registry with MLflow 3, automated retraining, and production monitoring. Throughout the week, participants work through large-scale problems drawn from real production practice (messy multi-source data, evolving schemas, models that degrade after deployment) with agentic AI assistants sharing the

implementation work at every stage: generating pipeline code, diagnosing failures, and refactoring under review, while participants build the judgment to direct and evaluate them. Participants finish able to design and operate the full path from raw data to a continuously improving model in production.

Learning Outcomes

- Design a lakehouse architecture on open table formats (Apache Iceberg, Delta Lake) that serves both analytics and machine learning workloads.
- Build reliable, performance-tuned data pipelines with Apache Spark 4 and layered SQL transformations with dbt.
- Orchestrate multi-stage data and ML workflows with Apache Airflow 3, including asset-based scheduling and backfills.
- Ingest and process streaming data with Apache Kafka 4 and Spark Structured Streaming.
- Enforce data quality with schema management, validation checks, and data contracts across batch and streaming sources.
- Build feature pipelines with point-in-time correctness and serve features consistently between training and inference.
- Scale model training beyond a single machine with efficient data loading and distributed training strategies.
- Track experiments and manage the model lifecycle with MLflow 3, from logged runs through registered, versioned models.
- Automate retraining and model promotion with validation gates, progressive deployment, and rollback strategies.
- Detect data drift, concept drift, and model degradation in production, and respond with structured incident practices.
- Apply agentic AI assistants across the platform: generating pipeline and training code, diagnosing failures, and reviewing changes for production safety.

Training Materials

Comprehensive courseware is distributed online at the start of class. All students receive a downloadable MP4 recording of the training.

Software Requirements

Students will need a computer capable of running Docker containers (16 GB of RAM recommended) to host the class data and ML stack. If students are

unable to configure a local environment, a cloud-based environment can be provided.

Training Topics

Production Data Architecture and the Lakehouse

- From warehouses and data lakes to the lakehouse
- Open table formats: Apache Iceberg and Delta Lake
- ACID transactions, time travel, and schema evolution on object storage
- Medallion architecture: bronze, silver, and gold layers
- Catalogs, governance, and table maintenance
- Where ML workloads fit in the platform

Large-Scale Processing with Apache Spark 4

- Spark architecture: driver, executors, and Spark Connect
- DataFrames, Spark SQL, and the VARIANT type for semi-structured data
- Joins, aggregations, and window functions at scale
- Partitioning, shuffles, and Adaptive Query Execution
- Reading and writing Iceberg and Delta tables from Spark
- Performance tuning and cost control

SQL Transformation with dbt

- dbt project structure: models, sources, and seeds
- Materializations: views, tables, and incremental models
- Jinja templating and macros
- Tests, documentation, and lineage
- Layered transformation design from staging to marts

Orchestration with Apache Airflow 3

- Airflow 3 architecture and the airflow.sdk authoring interface
- Authoring DAGs with the @dag and @task decorators
- Asset-based and event-driven scheduling
- Dynamic task mapping and deferrable operators
- Backfills, retries, and DAG versioning
- Orchestrating Spark and dbt jobs from Airflow

Streaming Ingestion with Apache Kafka 4

- Kafka 4 architecture: topics, partitions, and KRaft mode

- Producers, consumer groups, and delivery semantics
- Schema registries and message serialization
- Kafka Connect for source and sink integration
- Stream processing with Spark Structured Streaming
- Streaming data into lakehouse tables

Data Quality, Schemas, and Contracts

- Dimensions of data quality and where pipelines break
- Expectation-based validation in pipelines
- Schema evolution strategies across batch and streaming
- Data contracts between producers and consumers
- Freshness, completeness, and anomaly checks
- Data observability and lineage

Feature Engineering at Scale and Feature Stores

- Feature pipelines on Spark and dbt
- Point-in-time correctness and leakage prevention
- Offline and online feature stores
- Feature freshness, backfills, and reuse across teams
- Embeddings and streaming features

Distributed and Large-Scale Training

- When a single machine is no longer enough
- Data-parallel training and gradient synchronization
- Distributed training on CPU and GPU clusters
- Efficient training data loading from lakehouse tables
- Hyperparameter search at scale
- Checkpointing and fault tolerance for long-running jobs

Experiment Tracking and Model Registry with MLflow 3

- Experiment tracking: runs, parameters, metrics, and artifacts
- Logged models, signatures, and packaging in MLflow 3
- Model registry: versions, aliases, and approval workflows
- Model lineage from training data to deployed version
- Comparing and evaluating candidate models

Automated Retraining and Continuous Delivery

- CI/CD for machine learning vs. traditional software

To discuss this course and customizations:
Call: +1 434-509-6890 or Email: sales@cloudcontraptions.com

- Retraining triggers: schedules, data events, and drift signals
- Airflow-orchestrated retraining pipelines
- Validation gates and automated model promotion
- Shadow, canary, and blue-green deployment patterns
- Rollback and model version pinning

Production Monitoring, Drift, and Incident Response

- Monitoring predictions, latency, and business metrics
- Data drift and concept drift detection
- Delayed ground truth and proxy metrics
- Alerting thresholds and escalation for ML systems
- Root cause analysis across data and model layers
- Closing the loop: monitoring signals that drive retraining

Agentic AI Assistants Across the ML Platform

- Where agents help: pipeline code, tests, and migrations
- Generating and reviewing Spark, dbt, and DAG code with agents
- Agent-assisted debugging of data and training failures
- Review practices and guardrails for agent-written code
- Keeping human judgment in the loop for production changes

