

[DatabricksTraining.com](https://databrickstraining.com)

Azure Data Engineering

10-Week / 50-Topic Detailed Professional Syllabus

Course focus	PySpark + Azure + ADF + Databricks + Streaming + Fabric + AI/DevOps
Structure	10 weeks, 50 weekday topics, weekend masterclasses and 5 major projects
Learning style	Concepts + live coding + architecture + troubleshooting + project implementation

Course philosophy: build strong PySpark foundations first, then connect them to cloud-native data engineering, lakehouse design, streaming, orchestration, governance, testing, performance, CI/CD, and AI-assisted engineering productivity.

Course Overview

A detailed Azure Data Engineering curriculum that combines PySpark/Spark foundations with ADLS, ADF, Databricks, Delta Lake, Unity Catalog, streaming, Airflow, Microsoft Fabric, CI/CD and AI-assisted engineering.

Duration	10 weeks; 50 focused weekday topics plus hands-on projects and weekend masterclasses
Ideal for	Data engineers, ETL developers, Spark developers and professionals moving into Azure/Databricks data engineering
Delivery model	Instructor-led sessions + guided labs + projects + weekend masterclasses
Primary outcome	Ability to design, build, troubleshoot and explain production-oriented data pipelines.

What learners should be able to do

- Build ingestion and orchestration pipelines using ADF, ADLS and Databricks.
- Implement Bronze/Silver/Gold architectures with Delta Lake and Unity Catalog.
- Build streaming pipelines using Kafka/Event Hubs and Structured Streaming.
- Design metadata-driven ADF frameworks and CI/CD with Databricks Asset Bundles.
- Apply governance, security, testing, monitoring and Spark performance tuning.
- Understand where Microsoft Fabric complements or replaces parts of an Azure data platform.

Detailed Weekly Curriculum

Each numbered session below can be delivered as a focused class with demonstrations and coding exercises. Weekend sessions are used for AI-assisted engineering, metadata-driven design, architecture and project integration.

Week 1 - PySpark & Databricks Foundations | 10 Hours

1. PySpark & Databricks Introduction

- Spark architecture: driver, executors and cluster manager
- Jobs, stages, tasks and lazy evaluation
- Databricks workspace, notebooks and compute
- DataFrame-based engineering workflow

2. Databricks Date & Common Functions

- Date/time parsing
- String and numeric functions
- Conditional logic
- Arrays, maps and structs

3. Databricks Window Functions

- row_number, rank and dense_rank
- lag and lead
- Partitions and ordering
- Running totals and top-N use cases

4. Data Cleaning with Regular Expressions

- Null handling
- Duplicate handling
- Regex extraction/replacement
- Data standardization and validation

5. spark-submit Internals

- Submission lifecycle
- Driver and executor configuration
- Cores/memory/parallelism
- Reading logs and Spark UI evidence

Weekend Masterclass / Lab: Claude Code + Prompt Engineering Special Class

- Generate and review PySpark
- Debug Spark errors

- Prompt patterns for ETL/SQL
- AI-assisted coding with validation

Week 2 - Advanced Spark Data Processing | 10 Hours

6. Spark JSON Processing

- Nested JSON
- StructType
- Arrays/structs
- explode and schema handling

7. Spark JDBC Processing

- Oracle
- MySQL
- SQL Server
- Parallel JDBC reads and pushdown

8. XML, CSV & Pandas Data Processing

- XML patterns
- CSV options
- Pandas vs PySpark
- Pandas API on Spark overview

9. Parquet & ORC

- Columnar storage
- Compression
- Predicate pushdown
- Column pruning and partition design

10. Delta Lake vs Apache Iceberg

- ACID and table metadata
- Time travel and schema evolution
- Open-table format trade-offs
- When to choose Delta or Iceberg

Weekend Masterclass / Lab: Spark File-Format & Performance Lab

- Convert raw files to Parquet/Delta
- Inspect execution plans
- Compare scan behavior
- Practice partition-pruning scenarios

Week 3 - Azure Storage & Azure Data Factory | 10 Hours

11. Azure Introduction, Blob Storage & ADLS Gen2

- Subscriptions/resource groups
- Storage accounts and containers
- Blob vs ADLS Gen2
- Hierarchical namespace, access tiers and lake folder design

12. Azure Data Factory Introduction & Internals

- ADF architecture
- Pipelines, activities and datasets
- Linked services
- Integration Runtime basics

13. Core ADF Activities

- Copy Activity
- ForEach
- Lookup
- Get Metadata
- If Condition and dependency control

14. Databricks Notebook + Lookup + ForEach

- ADF to Databricks orchestration
- Passing parameters
- Dynamic notebook execution
- Loop-driven processing

15. Parameterization & Mapping Data Flows

- Pipeline parameters
- Dataset parameters
- Dynamic content/expression language
- Data Flow transformations and performance considerations

Weekend Masterclass / Lab: ADF with Claude Code - JSON-Driven Development

- Generate pipeline JSON safely
- Understand ADF ARM/resource structure
- Parameterize reusable pipelines
- Validate AI-generated JSON before deployment

Week 4 - ADF Advanced, Security & End-to-End Project | 10 Hours

16. Schedulers, Key Vault & Triggers

- Schedule/tumbling/event triggers
- Azure Key Vault integration
- Secret references
- Operational scheduling patterns

17. SCD Type 1, SCD Type 2 & Integration Runtime

- SCD concepts
- Mapping/SQL approaches
- Self-hosted IR
- Azure IR and network/connectivity considerations

18. Databricks + ADLS Integration

- ABFS access
- Managed identity/service principal concepts
- Unity Catalog storage credentials
- Secure data access

19. Data Governance & Security

- Azure RBAC
- Encryption
- Network security
- Managed identities, private endpoints and least privilege

20. ADF End-to-End Pipeline Project using Databricks

- Source ingestion
- ADF orchestration
- Databricks transformation
- ADLS Bronze/Silver/Gold
- Monitoring and rerun strategy

Weekend Masterclass / Lab: ADF Metadata-Driven Pipelines

- Configuration tables/files
- Dynamic source/target processing
- Reusable Copy/Notebook activities
- Centralized error handling and logging

Week 5 - Delta Lake, Medallion & Databricks Optimization | 10 Hours

21. Databricks Utilities

- dbutils.fs
- dbutils.secrets
- dbutils.widgets
- Parameter-driven notebook design

22. Delta Time Travel & Change Data Feed

- Version history
- Restore/read-as-of
- CDF for incremental downstream processing
- MERGE and audit use cases

23. Medallion Architecture on Azure

- Bronze/Silver/Gold responsibilities
- ADLS landing zones
- Data quality/quarantine
- Idempotent reprocessing

24. Spark Declarative Pipelines (SDP)

- Declarative pipeline concepts
- Dependency graphs
- Expectations/data quality
- Incremental processing and operational benefits

25. Databricks Optimization

- OPTIMIZE/VACUUM
- Partitioning and file sizing
- Z-Ordering and Liquid Clustering
- AQE, Photon, joins and skew

Weekend Masterclass / Lab: Lakehouse Federation + Delta Sharing

- Federated-query use cases
- Secure data sharing
- Internal/external consumers
- Governance considerations

Week 6 - Databricks Engineering, Deployment & Governance | 10 Hours

26. Databricks Asset Bundles (DABs)

- Bundle structure
- databricks.yml
- Environment targets
- Jobs/pipelines deployment

27. Databricks Lakeflow

- Lakeflow Connect overview
- Jobs and pipeline ecosystem
- Operational workflow patterns
- Ingestion-to-orchestration integration

28. Lakeflow Orchestration Advanced

- Task dependencies
- Parameters
- Retries and notifications
- Multi-task workflows and environment promotion

29. Unity Catalog Introduction & Auto CDC

- Metastore/catalog/schema/table hierarchy
- Managed vs external objects
- Lineage
- Auto CDC concepts for incremental changes

30. Unity Catalog Advanced

- Storage credentials/external locations
- Privileges
- Row filters and column masks
- Governed sharing and auditing

Weekend Masterclass / Lab: CI/CD with DABs

- Git-based development
- Dev/Test/Prod targets
- GitHub Actions pipeline concept
- Deployment validation and rollback discussion

Week 7 - Streaming on Azure | 10 Hours

31. Spark Structured Streaming Introduction

- Streaming DataFrames
- Micro-batches

- Output modes
- State and checkpoints

32. Kafka with Spark

- Producer/consumer basics
- Topics/partitions/offsets
- Spark Kafka connector
- Replay and failure handling

33. NiFi + Kafka + Spark Streaming

- NiFi ingestion patterns
- Kafka transport layer
- Spark consumption
- End-to-end observability points

34. Azure Event Hubs Streaming Project

- Event Hubs architecture
- Partitions and consumer groups
- Kafka-compatible endpoint concepts
- Databricks Structured Streaming integration

35. Structured Streaming, Kafka & Auto Loader Internals

- Checkpoints
- Watermarks
- Exactly-once-oriented design
- Auto Loader schema inference/evolution

Weekend Masterclass / Lab: Real-Time Azure Streaming Project

- Producer/Event Hub -> Databricks
- Bronze streaming table
- Silver transformations
- Gold aggregates + checkpoint/replay demonstration

Week 8 - Airflow, Testing, Observability & Spark Performance | 10 Hours

36. Apache Airflow Introduction

- DAGs, tasks and operators
- Scheduler/executor
- Variables/connections
- XCom basics

37. Airflow Advanced Orchestration

- TaskFlow API
- Sensors and trigger rules
- Dynamic task mapping
- Databricks/ADF integration patterns

38. Testing & Observability

- PySpark unit tests
- Data-quality tests
- Pipeline logging and metrics
- Databricks/Azure Monitor concepts

39. Performance Tuning Deep Dive

- Explain plans
- Partitioning
- Join strategies
- Memory, shuffle and skew optimization

40. Spark RDD Internals

- RDD lineage
- Narrow vs wide transformations
- DAG construction
- Fault tolerance and recomputation

Weekend Masterclass / Lab: Production Troubleshooting Workshop

- Diagnose a slow Spark job
- Trace pipeline failure across ADF/Databricks
- Analyze logs and Spark UI
- Prioritize performance fixes

Week 9 - Microsoft Fabric & OneLake | 10 Hours

41. Microsoft Fabric & OneLake Core Architecture

- Fabric workloads
- OneLake concepts
- Workspaces and capacities
- How Fabric fits with Azure data platforms

42. Fabric Lakehouse vs Synapse Data Warehouse

- Lakehouse vs warehouse
- Delta tables

- SQL analytics endpoint
- When to choose each architecture

43. Real-Time Intelligence & Data Activator

- Real-time ingestion/analytics concepts
- Eventstreams/KQL concepts
- Triggering actions from data changes
- Operational scenarios

44. Semantic Models in Microsoft Fabric

- Model relationships
- Measures and star schema thinking
- Direct Lake concepts
- Performance and governance considerations

45. Prepare AI-Ready Analytics Data in Fabric

- Curated data products
- Semantic consistency
- Security/governance
- Preparing trusted data for AI/analytics consumers

Weekend Masterclass / Lab: Azure vs Fabric Architecture Workshop

- ADF/Databricks/Fabric role comparison
- Lakehouse vs warehouse selection
- Migration and coexistence patterns
- Cost and operating-model discussion

Week 10 - AI-Assisted Engineering, CI/CD & Interview Preparation | 10 Hours

46. Claude Code for Data Engineers

- Repository-aware coding
- Generate/debug PySpark
- Generate ADF/Lakeflow examples
- Testing AI output

47. Prompt Engineering

- Context engineering
- Structured prompts
- Debugging prompts
- Architecture prompts

48. Codex & GitHub Copilot for Productivity

- Code generation
- Unit tests
- Documentation
- Refactoring and review

49. GitHub Actions CI/CD

- Git workflow
- Build/test
- Databricks Asset Bundle deployment
- Environment/secrets management

50. Interview Tips & Resume Preparation

- PySpark and SQL scenarios
- ADF/Databricks architecture
- Streaming/performance troubleshooting
- Resume and project explanation

Weekend Masterclass / Lab: Mock Azure Data Engineer Architecture & Interview Clinic

- Design an end-to-end Azure platform
- Explain security and governance choices
- Practice troubleshooting questions
- Convert course projects into resume-ready experience statements

Hands-On Projects

Projects are deliberately aligned with the course sequence so students first learn a concept, then implement it as part of a realistic pipeline.

Project 1 - ADF + Databricks Batch Pipeline

Source DB / Files -> ADF -> ADLS Bronze -> Databricks -> Silver/Gold Delta -> Analytics

- Linked services/datasets
- Parameterization
- Databricks orchestration
- Medallion architecture

Project 2 - Metadata-Driven ADF Framework

Metadata Config -> ADF Lookup/ForEach -> Dynamic Copy/Notebook -> Logging -> Target

- Reusable pipeline design
- Dynamic expressions
- Centralized configuration
- Operational error handling

Project 3 - Azure Databricks Lakehouse

ADLS Landing -> Auto Loader -> Bronze -> Silver -> Gold -> Unity Catalog -> BI/Consumers

- Incremental ingestion
- Delta Lake
- Data quality
- Governance and optimization

Project 4 - Azure Real-Time Streaming

Producer / Kafka -> Azure Event Hubs -> Databricks Structured Streaming -> Delta -> Gold

- Event Hubs
- Kafka compatibility
- Checkpointing
- Watermarks and late-data handling

Project 5 - CI/CD & Production Deployment

Git -> GitHub Actions -> Databricks Asset Bundles -> Dev/Test/Prod -> Validation

- Version control
- Automated tests
- Environment promotion
- Deployment governance

Technology Stack Covered

Area	Technologies / Concepts
Core Engineering	PySpark, Spark SQL, JDBC, JSON, XML, CSV, Parquet, ORC, Delta Lake, Iceberg concepts
Azure Storage	Azure Blob Storage, ADLS Gen2
Azure Integration	Azure Data Factory, Integration Runtime, Triggers, Mapping Data Flows, Key Vault
Azure Streaming	Azure Event Hubs, Apache Kafka, Structured Streaming, Auto Loader, NiFi
Databricks	Databricks, Delta Lake, Unity Catalog, Lakeflow, Spark Declarative Pipelines, DABs
Governance & Security	Azure RBAC, Managed Identity, Service Principal concepts, Encryption, Private networking, Unity Catalog
Orchestration & Ops	ADF, Airflow, Databricks Workflows/Lakeflow, Testing, Monitoring, Observability
Microsoft Fabric	OneLake, Fabric Lakehouse, Warehouse, Real-Time Intelligence, Semantic Models
DevOps & AI	Git, GitHub, GitHub Actions, Claude Code, Codex, GitHub Copilot, Prompt Engineering

Note: The curriculum focuses on data-engineering use cases. General cloud administration topics are covered only when they directly affect pipeline design, security, performance or operations.