Mastering Apache Airflow

Workflow Orchestration for Data Engineering Excellence

Technologies: Apache Airflow • AWS MWAA • GCP Composer • Azure ADF • Kubernetes • Spark • Snowflake • dbt • Kafka • Docker • CI/CD

databrickstraining.in | Training & Placement Excellence
Phone: +91-8500002025 | **Email:** info@databrickstraining.in

★ ★ ★

COURSE OVERVIEW
This comprehensive 110-hour course covers Apache Airflow from fundamentals to enterprise-grade production deployments. You will master DAG authoring, all major operators, multi-cloud integrations (AWS, Azure, GCP), Spark orchestration, Snowflake & dbt pipelines, and production Kubernetes deployments. Ideal for Data Engineers, DevOps Engineers, and Analytics Engineers who need to schedule, monitor, and manage complex data pipelines at scale.

Who Should Attend? Data Engineers, Pipeline Developers, Analytics Engineers, Cloud Engineers, DevOps practitioners building data infrastructure.

Prerequisites: Basic Python, familiarity with SQL, and introductory cloud concepts (AWS/Azure/GCP).

Training Format: Instructor-led sessions with hands-on labs, real-world projects, mock interviews, and certification preparation.

Module 1: Workflow Orchestration & Airflow Fundamentals | Duration: 3 Hours | Level: Beginner

What is Workflow Orchestration & Why It Matters
Problems Without Orchestration — Manual Pipeline Pitfalls
Orchestration Tools Comparison
• Apache Airflow vs Luigi vs Prefect vs Dagster
• Airflow vs dbt Cloud vs AWS Step Functions
• Decision Guide — When to Use Airflow
What is Apache Airflow & Why It Is the Industry Standard
Airflow History — Airbnb to Apache Foundation
Airflow Use Cases — Real Industry Examples
• Data pipeline scheduling, ETL/ELT orchestration
• ML pipeline automation, Multi-cloud pipeline management
• Data quality monitoring
Airflow Architecture — Complete Overview
• Web Server — UI & REST API
• Scheduler — Trigger & Monitor DAGs
• Executor — Run Tasks | Worker — Execute Task Code
• Metadata Database — State & History
• Message Broker — Celery/Redis Queue
Airflow 1.x vs Airflow 2.x — Key Differences
Airflow Versions — 2.6, 2.7, 2.8, 2.9, 2.10
Labs: Explore Airflow Architecture Diagrams, Set Up Airflow Locally with Docker Compose

Module 2: Airflow Installation & Setup | Duration: 3 Hours | Level: Beginner

Airflow Installation Methods
• pip install — Local Development
• Docker Compose — Most Common for Learning
• Helm Chart — Kubernetes Production
• Managed Airflow — AWS MWAA, GCP Composer, Astronomer
Docker Compose Setup — Step by Step
• docker-compose.yaml, Start/Stop Airflow, Access UI on localhost:8080
Airflow Configuration — airflow.cfg
• Core settings, Database URI, SMTP, Webserver auth
Airflow Environment Variables — Override Config
Airflow Connections — Connect to External Systems
• Create via UI, CLI, Environment Variables, Secrets Backend
Airflow Variables — Store Global Config Values
Airflow Users & Roles — RBAC Setup
Airflow Pools — Limit Concurrent Tasks
Airflow CLI — Essential Commands
• dags list, tasks test, dags trigger, backfill, db init & upgrade
Labs: Docker Compose Airflow Full Setup, Create Connections & Variables via UI/CLI
Quiz: Airflow Setup — 15 Questions

Module 3: DAGs — Directed Acyclic Graphs | Duration: 5 Hours | Level: Intermediate

What is a DAG — Directed Acyclic Graph
DAG Anatomy — Complete Breakdown
• dag_id, schedule_interval, start_date, catchup, tags
• default_args, max_active_runs, max_active_tasks
DAG Scheduling
• Cron expressions, Cron presets (@daily, @hourly, @weekly)
• Timedelta, Timetables — Custom scheduling
• Data-Aware Scheduling — Datasets
Logical Date vs Execution Date vs Data Interval
Catchup — Backfill Historical Runs
DAG & Task Run States — Running, Success, Failed, Skipped
Task Dependencies
• set_upstream() & set_downstream()
• >> and << bitshift operators
• chain(), cross_downstream()
Trigger Rules: all_success, all_failed, all_done, one_success, none_failed, none_skipped
DAG Params — Pass Runtime Parameters
Context Manager DAGs (with DAG()) vs Decorator DAGs (@dag & @task)
Labs: First DAG, Complex Task Graph, DAG Params Configuration
Quiz: DAGs & Scheduling — 20 Questions

Module 4: Airflow Operators — Complete Guide | Duration: 6 Hours | Level: Intermediate

SECTION A — Core Operators

• PythonOperator — python_callable, op_args, op_kwargs, templates_dict
• BashOperator — bash_command, env variables, append_env
• PythonVirtualenvOperator & ExternalPythonOperator
• BranchPythonOperator — Conditional Branching
• ShortCircuitOperator — Skip Downstream on False
• EmailOperator, DummyOperator, LatestOnlyOperator, TriggerDagRunOperator

SECTION B — File & HTTP Operators

• SimpleHttpOperator, FileSensor, HttpSensor, PythonSensor

SECTION C — Database Operators

• PostgresOperator, MySQLOperator, MsSqlOperator, SqliteOperator, GenericTransfer

SECTION D — Taskflow API — Modern Airflow Style

• @task Decorator, XCom with Taskflow (Automatic Push & Pull)
• @task.branch, @task.sensor, @task.virtualenv
• Multiple Outputs — dict return, Taskflow with Traditional Operators

SECTION E — Sensors

• Sensor Modes: Poke vs Reschedule
• poke_interval, timeout, soft_fail
• FileSensor, HttpSensor, S3KeySensor, ExternalTaskSensor
• DateTimeSensor, TimeDeltaSensor, Custom Sensor

SECTION F — Best Practices

• Idempotent & Atomic tasks, XCom size limits, Template Fields

Labs: BranchPythonOperator, Sensor Pipeline, Taskflow API DAG, REST API Pipeline

Quiz: Operators — 20 Questions

Module 5: XComs & Task Communication | Duration: 3 Hours | Level: Intermediate

What is XCom — Cross-Communication Between Tasks
XCom Push — xcom_push() & return value
XCom Pull — xcom_pull() with task_id & key
XCom in Jinja Templates
XCom Size Limits — What NOT to Pass
XCom with Taskflow API — Automatic XCom
XCom Backends — Custom Storage
• Default — Metadata DB (small data only)
• S3 XCom Backend, GCS XCom Backend, ADLS XCom Backend
Multiple XCom Outputs — Return Dict
XCom in BranchPythonOperator

Labs: Pass Data Between Tasks with XCom, S3 XCom Backend for Large Payloads

Module 6: Airflow Executors | Duration: 3 Hours | Level: Intermediate

What is an Executor — How Tasks Run
Executor Types
• SequentialExecutor — One task at a time (SQLite only)
• LocalExecutor — Parallel on single machine (PostgreSQL)

• CeleryExecutor — Distributed, multiple workers
• KubernetesExecutor — Each task in a Kubernetes pod
• CeleryKubernetesExecutor & LocalKubernetesExecutor — Hybrid
LocalExecutor — Best for Medium Workloads
• parallelism, max_active_tasks_per_dag
CeleryExecutor — Production Standard
• Redis or RabbitMQ as broker, Multiple worker machines
• Flower — Monitor Celery workers
KubernetesExecutor — Cloud-Native Production
• Each task = separate Kubernetes pod, No idle workers
• Pod templates — resource requests & limits
Executor Comparison — When to Use Each
Labs: LocalExecutor Parallel Tasks, CeleryExecutor with Redis
Quiz: Executors — 10 Questions

Module 7: Airflow Templates & Jinja | Duration: 3 Hours | Level: Intermediate

What is Jinja Templating in Airflow

Template Fields — Which Fields Support Jinja

Airflow Macros — Built-In Template Variables

• {{ ds }} — Execution date YYYY-MM-DD

• {{ ds_nodash }}, {{ ts }}, {{ prev_ds }}, {{ next_ds }}

• {{ dag.dag_id }}, {{ task.task_id }}, {{ run_id }}

• {{ params }}, {{ var.value.my_var }}, {{ conn.my_conn.host }}

Custom Macros — user_defined_macros

Jinja Filters — Format Dates & Strings

Rendering Templates — Rendered Template UI

Template Files — SQL & Scripts in Templates

Dynamic SQL with Jinja

Labs: Dynamic File Paths with Jinja, Parameterised SQL Queries

Module 8: Airflow AWS Integration | Duration: 6 Hours | Level: Intermediate-Advanced

apache-airflow-providers-amazon Setup & AWS Connection (Access Key, IAM Role)

Amazon S3 Operators

• S3CreateBucketOperator, S3DeleteBucketOperator, S3FileTransformOperator

• S3ListOperator, S3CopyObjectOperator, S3KeySensor

• S3ToLocalFilesystemOperator, LocalFilesystemToS3Operator, S3ToSnowflakeOperator

AWS Glue Operators

• GlueJobOperator, GlueJobSensor, GlueCrawlerOperator, GlueCatalogOperator

AWS EMR Operators

• EmrCreateJobFlowOperator, EmrAddStepsOperator, EmrStepSensor

• EmrTerminateJobFlowOperator, EmrServerlessStartJobRunOperator

AWS Lambda & Redshift Operators

• LambdaInvokeFunctionOperator, RedshiftSQLOperator

• S3ToRedshiftOperator, RedshiftToS3Operator

AWS Step Functions, SNS, SQS & Athena

• StepFunctionStartExecutionOperator, SnsPublishOperator

• SqsSensor, AthenaOperator, AthenaSensor

Labs: S3 → Redshift Pipeline, EMR Spark Job, Glue ETL, Full AWS Pipeline with SNS Alert

Quiz: Airflow AWS — 15 Questions

Module 9: Airflow Azure Integration | Duration: 5 Hours | Level: Intermediate-Advanced

apache-airflow-providers-microsoft-azure Setup (Service Principal, Managed Identity)

Azure Data Lake Storage (ADLS)

• ADLSCreateDirectoryOperator, ADLSDeleteOperator, ADLSListOperator

• LocalToADLSOperator, ADLSToLocalOperator

Azure Data Factory (ADF)

• AzureDataFactoryRunPipelineOperator — Trigger ADF Pipeline

• AzureDataFactoryPipelineRunSensor, AzureDataFactoryGetPipelineRunOperator

Azure Blob Storage

• WasbHook, WasbBlobSensor

Azure SQL & Synapse

• MsSqlOperator, AzureSynapseRunSparkBatchOperator, AzureSynapseSensor

Azure Databricks Operators

• DatabricksRunNowOperator — Trigger Databricks Job

• DatabricksSubmitRunOperator — Submit Notebook/JAR

• DatabricksTaskOperator, DatabricksSensor, DatabricksCopyIntoOperator

Azure Container & Kubernetes

• AzureContainerInstanceOperator, KubernetesPodOperator on AKS

Labs: ADLS + ADF Pipeline, Azure Databricks Notebook, Full Azure → Synapse Pipeline

Quiz: Airflow Azure — 15 Questions

Module 10: Airflow GCP Integration | Duration: 5 Hours | Level: Intermediate-Advanced

apache-airflow-providers-google Setup (Service Account, Workload Identity)

Google Cloud Storage (GCS)

• GCSCreateBucketOperator, GCSListObjectsOperator, GCSToGCSOperator

• LocalFilesystemToGCSOperator, GCSObjectExistenceSensor

BigQuery Operators

• BigQueryInsertJobOperator — Run BQ Query

• BigQueryCheckOperator, BigQueryValueCheckOperator

• BigQueryToGCSOperator, GCSToBigQueryOperator, BigQueryTableExistenceSensor

Dataflow Operators

• DataflowCreatePythonJobOperator, DataflowStartFlexTemplateOperator, DataflowJobSensor

Dataproc Operators

• DataprocCreateClusterOperator, DataprocSubmitJobOperator

• DataprocDeleteClusterOperator, DataprocCreateBatchOperator (Serverless Spark)

Cloud Functions, Run, PubSub

• CloudFunctionDeployFunctionOperator, CloudRunExecuteJobOperator

• PubSubPublishMessageOperator, PubSubPullOperator, PubSubSensor

Labs: GCS → BigQuery Load, Dataproc PySpark Job, Full GCP → BQ → Looker Pipeline

Quiz: Airflow GCP — 15 Questions

Module 11: Airflow Spark Integration | Duration: 5 Hours | Level: Advanced

Spark + Airflow — Orchestration Patterns

SparkSubmitOperator — Submit Spark Applications

• application, conn_id, conf, executor_cores, executor_memory, num_executors

• jars, py_files, verbose

SparkSqlOperator, SparkJDBCOperator

Spark on YARN — Submit via Airflow

Spark on Kubernetes — KubernetesPodOperator

Spark on AWS EMR

• EmrAddStepsOperator with Spark step + EmrStepSensor

Spark on GCP Dataproc

• DataprocSubmitJobOperator — PySpark job

Spark on Azure Databricks

• DatabricksSubmitRunOperator, DatabricksRunNowOperator

Livy — REST API for Spark Jobs

• LivyOperator — Submit Spark via REST, LivySensor

Spark Structured Streaming — Trigger from Airflow

Passing Parameters to Spark from Airflow (--conf, app args, Jinja)

Monitor Spark Jobs from Airflow (Spark UI links, success/failure detection)

Labs: SparkSubmitOperator on YARN, EMR Spark Job, Dataproc GCP, Databricks Azure

Quiz: Spark + Airflow — 15 Questions

Module 12: Airflow Snowflake Integration | Duration: 4 Hours | Level: Advanced

apache-airflow-providers-snowflake Setup & Connection (Key Pair, OAuth)

Snowflake Operators

• SnowflakeOperator — Run SQL (warehouse, database, schema override)
• SnowflakeCheckOperator, SnowflakeValueCheckOperator
• SnowflakeIntervalCheckOperator, SnowflakeAsyncOperator
• SnowflakeSensor, S3ToSnowflakeOperator, SnowflakeToSlackOperator

Snowflake Pipeline Patterns with Airflow

• Pattern 1: Daily ELT — Extract → S3 → Snowpipe → dbt → Validate
• Pattern 2: Incremental Load — MERGE + Jinja date partitioning
• Pattern 3: Snowflake + dbt Orchestration — Cosmos / BashOperator
• Pattern 4: Data Quality Gate — CheckOperator + BranchPythonOperator

Snowflake Streams + Tasks vs Airflow — Decision Guide

Labs: Snowflake ELT Full Orchestration, S3 → Snowflake → dbt → DQ Check
• Snowflake + dbt Cloud — Trigger via Airflow Admin API

Quiz: Snowflake + Airflow — 15 Questions

Module 13: Airflow dbt Integration | Duration: 4 Hours | Level: Advanced

Why Orchestrate dbt with Airflow
dbt Core + Airflow — BashOperator Approach
• Run dbt commands: dbt run, dbt test, dbt source freshness
• Parse dbt results in Airflow
astronomer-cosmos — Native dbt + Airflow
• Auto-generate DAGs from dbt project
• Each dbt model = Airflow task, Task groups per folder
• Run dbt tests as Airflow tasks
• Selective model runs — tags & selectors
dbt Cloud Operators
• DbtCloudRunJobOperator, DbtCloudJobRunSensor
• DbtCloudGetJobRunArtifactOperator
Airflow + dbt Architecture Patterns
• BashOperator (Simple), Cosmos (Full DAG generation), dbt Cloud API (Managed)
Passing Airflow context to dbt — vars, date macros
dbt Test Results → Airflow Alerting
Labs: Cosmos Auto-Generate DAG, dbt Cloud + Airflow Pipeline, Selective Model Runs

Module 14: Airflow Database & Warehouse Operators | Duration: 4 Hours | Level: Intermediate-Advanced

PostgreSQL Operators: PostgresOperator, PostgresToS3Operator, S3ToPostgresOperator
MySQL Operators: MySQLOperator, MySQLToS3Operator, S3ToMySQLOperator
Microsoft SQL Server: MsSqlOperator, MsSqlToHiveOperator
Hive Operators: HiveOperator, HiveToMySqlOperator, HiveToDynamoDBOperator
SparkSQLOperator, PrestoOperator, TrinoOperator
Databricks Operators — Full List
• DatabricksRunNowOperator, DatabricksSubmitRunOperator
• DatabricksTaskOperator, DatabricksCopyIntoOperator, DatabricksSqlOperator
Generic Transfer Pattern — SQLToSQLOperator (Any DB to Any DB)
Labs: Postgres → S3 → Snowflake Full Pipeline, MySQL → Databricks Cross-Platform

Module 15: Airflow Advanced DAG Patterns | Duration: 5 Hours | Level: Advanced

Dynamic DAGs — Generate DAGs Programmatically
• Loop over config to create multiple DAGs
• YAML/JSON-driven DAG generation
• Database-driven DAG factory
Dynamic Task Mapping — Parallel Tasks at Runtime
• .expand() — Map over list of inputs
• .expand_kwargs() — Map over list of dicts
• .partial() — Fixed + dynamic arguments
SubDAGs — Deprecated but Important to Know
TaskGroups — Visually Group Tasks
• Create TaskGroup, Nested TaskGroups, TaskGroup dependencies
DAG Dependencies — Cross-DAG Orchestration
• TriggerDagRunOperator, ExternalTaskSensor, Datasets

Datasets — Event-Driven Scheduling

• Define Dataset (inlet & outlet), Schedule DAG on Dataset update

Conditional Logic: BranchPythonOperator, ShortCircuitOperator, @task.branch

Callback Functions: on_success_callback, on_failure_callback, on_retry_callback

SLA — Service Level Agreements & sla_miss_callback

Labs: Dynamic Task Mapping (100 Files), YAML DAG Factory, Datasets Event Pipeline, TaskGroups

Quiz: Advanced DAG Patterns — 20 Questions

Module 16: Airflow Production Deployment | Duration: 5 Hours | Level: Advanced

Production Airflow Architecture
• HA Webserver, Multiple Schedulers, CeleryExecutor + Redis, PostgreSQL DB
Airflow on Kubernetes — Helm Chart
• Install with Helm, values.yaml configuration
• KubernetesExecutor in production, Pod templates, Persistent volumes, Ingress
Managed Airflow Services
• AWS MWAA — Setup, S3 DAG deployment, environment sizing, VPC & private access
• GCP Cloud Composer — Composer 2 Autopilot, GCS DAG deployment, scaling
• Astronomer — Enterprise Airflow, Astro CLI, Deployment environments
DAG Deployment Strategies
• Git-Sync, CI/CD (GitHub Actions), S3 DAG Bucket
Airflow Logging
• Local, S3, GCS, Azure Blob, Elasticsearch logging
Airflow Metrics — StatsD & Prometheus
Grafana Dashboard for Airflow Metrics
Labs: Airflow on Kubernetes (Helm), AWS MWAA Deploy, GCP Composer DAG, CI/CD Deploy

Module 17: Airflow Security & Governance | Duration: 3 Hours | Level: Advanced

Airflow RBAC — Role-Based Access Control
• Built-in roles: Admin, Op, User, Viewer, Public
• Custom roles, DAG-level access
Airflow Authentication Providers
• Username & Password, OAuth (Google, GitHub, Okta)
• LDAP & Active Directory, SAML Enterprise SSO
Secrets Backend — Secure Credentials
• AWS Secrets Manager, GCP Secret Manager
• Azure Key Vault, HashiCorp Vault
Airflow Connection Encryption — Fernet Key
Network Security — Airflow Behind Firewall
DAG Code Visibility — Hide Code from UI
Audit Logs — Track User Actions
Airflow with VPC — Private Deployment
Labs: OAuth Setup (Google Login), AWS Secrets Manager Secure Connections

Module 18: Airflow Monitoring & Alerting | Duration: 4 Hours | Level: Advanced

Airflow UI Monitoring
• DAG view, Grid view, Graph view, Gantt chart, Task logs, Audit logs
Airflow Metrics — What to Monitor
• DAG success rate, Task duration trends, Scheduler heartbeat
• Task queue depth, Zombie tasks
StatsD — Airflow Metrics Collection
Prometheus + Grafana — Dashboard
• Airflow exporter for Prometheus, Grafana dashboard templates

Alerting Strategies

• Email alerts (SMTP setup)

• Slack alerts — SlackWebhookOperator

• PagerDuty — On-call alerting

• Microsoft Teams — TeamsWebhookOperator

Callback-Based Alerting: on_failure_callback, on_retry_callback, SLA miss callback

Dead Letter Pattern — Handle Repeated Failures

Airflow REST API — Monitor Programmatically

• Get DAG run status, Trigger DAGs, External monitoring integration

Labs: Slack Failure Notifications, Grafana Health Dashboard, REST API Monitoring

Module 19: Airflow CI/CD & Best Practices | Duration: 4 Hours | Level: Advanced

DAG Development Best Practices

• Idempotent & Atomic tasks, Avoid heavy top-level code

• Use XCom for small data only, Naming conventions, Default args (DRY)

DAG Testing

• Unit test DAG structure (pytest), Test task logic in isolation

• Integration test — airflow tasks test command

• DAG validation — check for cycles & import errors

CI/CD for Airflow DAGs

• GitHub Actions — Lint with flake8 & black

• Run pytest on DAG tests, Deploy to S3/GCS on merge

• Pre-commit hooks — Auto-format & lint

Git Workflow for DAG Development

• Feature branches, PR review, Environment branches (dev, staging, prod)

Astro CLI — Local Development & Testing

• astro dev start, astro dev pytest, astro deploy

Common DAG Mistakes — Avoid These

• Top-level DB calls, Dynamic dates in start_date

• Mutable default args dict, Missing catchup=False

Labs: pytest DAG Full Test Suite, GitHub Actions CI/CD, Astro CLI Workflow

Module 20: End-to-End Airflow Projects | Duration: 10 Hours | Level: Advanced

Project 1 — Multi-Cloud Data Platform (AWS)

• S3KeySensor → EMR PySpark → S3ToSnowflakeOperator → dbt → SnowflakeCheckOperator → Slack

• CI/CD: GitHub Actions → S3 DAG deployment

Project 2 — Azure Data Engineering Pipeline

• Azure SQL → ADF Copy → DatabricksRunNowOperator → Databricks Delta MERGE → Synapse

• EmailOperator on failure, daily schedule with retry logic

Project 3 — GCP Data Engineering Pipeline

• PostgreSQL → GCS → Dataproc PySpark → BigQueryInsertJobOperator → BigQueryCheckOperator

• PubSubPublishMessageOperator notify, deployed on Cloud Composer

Project 4 — Snowflake + dbt Orchestration Pipeline

• Salesforce, MySQL, S3 → Fivetran → Cosmos DAG → dbt tests → SnowflakeCheckOperator

• dbt Snapshot SCD Type 2, SlackWebhookOperator on test failures

Project 5 — Spark + Airflow + Snowflake Pipeline

• Kafka → S3 → SparkSubmitOperator → S3ToSnowflakeOperator → SnowflakeOperator

• Row count & freshness checks, Slack on failure, Email on SLA miss

• Kubernetes (Helm) deployment + GitHub Actions CI/CD

Module 21: Certification & Interview Preparation | Duration: 3 Hours | Level: Advanced

Apache Airflow Certification — Overview

• Astronomer Certification for Apache Airflow

• Exam format, question types, passing score

Exam Domain Coverage

• DAG authoring — scheduling, dependencies

• Operators & hooks, XCom & task communication

• Executors & scaling, Security & authentication

• Monitoring & alerting

Cloud Certification Airflow Topics

• AWS Certified Data Engineer — MWAA topics

• GCP Professional Data Engineer — Composer topics

• Azure Data Engineer DP-203 — ADF vs Airflow

Top 60 Apache Airflow Interview Questions & Answers

• Architecture design, DAG design scenarios

• Operator selection, Performance & scaling

• Troubleshooting, Real-world pipeline design

Career & Salary

• Resume Building for Data Engineering Roles

• LinkedIn Profile Optimization

• Salary Negotiation — ■8 LPA to ■35 LPA (India)

• Global market rates: US $110K-170K, EU €70K-110K

BONUS TOPICS: Advanced Airflow Mastery

1. Custom Operators & Hooks
• Build custom operators from BaseOperator
• Build custom hooks from BaseHook
• Publish to PyPI — share across teams
• Operator plugins — extend Airflow UI

2. Airflow on Kubernetes — Deep Dive
• KubernetesPodOperator — full config guide
• Pod template files — custom resource limits
• Sidecar containers in pod operator
• Kubernetes secrets in pod operator
• GPU pods for ML workloads

3. Airflow Performance Tuning
• Scheduler performance settings
• Parallelism & concurrency tuning
• DAG file processing interval
• Zombie task detection & cleanup
• Metadata DB optimization (PostgreSQL tuning)

4. Airflow + ML Pipelines
• Feature engineering DAGs
• Model training orchestration
• MLflow integration with Airflow
• Model deployment pipelines
• A/B testing DAG patterns

5. Airflow + Kafka (Event-Driven)
• Trigger DAGs from Kafka messages
• KafkaConsumerOperator & Sensor
• Event-driven pipeline patterns
• Combining Kafka + Datasets scheduling

6. Airflow REST API — Automation
• Trigger DAGs programmatically
• Get DAG run status from external systems
• Airflow API authentication (JWT)
• Build monitoring dashboards using API

7. Multi-Tenant Airflow
• Multiple team DAG isolation
• DAG-level permissions per team
• Resource pool per team
• Namespace isolation on Kubernetes

8. Airflow + dbt Mesh
• Orchestrate multiple dbt projects
• Cross-project dependency in Airflow
• Producer/Consumer DAG pattern
• Selective model run strategies

9. Advanced Retry & Error Handling

• Exponential backoff retry strategies
• Dead letter queue pattern
• Custom failure recovery DAGs
• Automatic rerun on partial failure

10. Airflow Cost Optimization
• Right-size worker instances
• KubernetesExecutor vs CeleryExecutor cost
• MWAA vs Composer vs Astronomer pricing
• Optimize DAG scheduling frequency

AIRFLOW INTERVIEW PREPARATION & CAREER TIPS

1. Technical Interview Strategy
• Know Airflow architecture inside out — every component
• Explain DAG design decisions with trade-offs
• Discuss scaling — CeleryExecutor vs KubernetesExecutor
• Show real experience with cloud integrations (AWS/Azure/GCP)

2. Top Interview Questions
• "What happens when the Airflow Scheduler restarts mid-run?"
• "How would you design a DAG that processes 1000 files in parallel?"
• "Explain the difference between Poke and Reschedule sensor modes"
• "How do you handle secrets in Airflow?"
• "What is the difference between ExternalTaskSensor and Datasets?"
• "How do you test Airflow DAGs before deploying to production?"
• "How would you implement SCD Type 2 in an Airflow DAG?"

3. Resume Bullet Points
■ Built Airflow pipelines
■ Designed 50+ production Airflow DAGs orchestrating PySpark, dbt, and Snowflake pipelines, reducing manual intervention by 90%

■ Used Airflow on AWS
■ Deployed Apache Airflow on AWS MWAA with KubernetesExecutor, supporting 200+ daily DAG runs with 99.9% uptime

4. Skills Checklist for Job Applications
■ DAG authoring (classic + Taskflow API)
■ At least 2 cloud provider integrations
■ Spark orchestration via Airflow
■ dbt + Airflow integration
■ Production deployment (MWAA/Composer/Kubernetes)
■ CI/CD for DAGs
■ Monitoring & alerting setup

5. Certifications Roadmap
Month 1: Astronomer Certification for Apache Airflow
Month 2-3: AWS Certified Data Engineer / GCP Professional Data Engineer
Month 4-5: Azure DP-203 Data Engineering
Month 6+: Databricks Certified Data Engineer

6. Salary Expectations (India)
• Entry level (0-2 years): ■8-12 LPA
• Mid-level (2-5 years): ■15-25 LPA
• Senior (5+ years): ■25-35 LPA
• Lead/Architect: ■35-55+ LPA

7. Salary Expectations (Global)
• US: $110K-175K base
• Canada: CAD $100K-145K
• EU/UK: €75K-120K
• Remote roles (India-based): 1.3-1.5x Indian market

RESUME & CAREER SKILL UPGRADE GUIDE

WHAT HIRING MANAGERS LOOK FOR

Must-Have Technical Skills

■ Python (pandas, PySpark, Airflow DAG authoring)
■ SQL (PostgreSQL, Snowflake, BigQuery)
■ Apache Airflow (DAGs, Operators, Sensors, Executors)
■ Cloud (AWS/Azure/GCP) data services
■ Spark orchestration via Airflow
■ Docker & containerization basics
■ Git & version control

Nice-to-Have Skills

■ Kubernetes (KubernetesPodOperator, Helm)
■ dbt + Cosmos integration
■ CI/CD (GitHub Actions, GitLab CI)
■ Terraform (infrastructure as code)
■ Kafka (event-driven pipelines)
■ MLflow (ML pipeline orchestration)

SKILLS PROGRESSION PATH

Year 1: Python + SQL + Airflow basics (DAGs, Operators)
Year 2: Cloud integrations + Spark + dbt + Production deployment
Year 3: Kubernetes + Security + CI/CD + Monitoring
Year 4+: Architecture design + Multi-cloud + Leadership

PROJECT PORTFOLIO CHECKLIST

■ End-to-end data pipeline (ingest → transform → serve)
■ Multi-cloud integration DAG (at least 2 clouds)
■ Dynamic Task Mapping DAG (parallel processing)
■ Spark orchestration via Airflow
■ dbt + Airflow (Cosmos or BashOperator)
■ Production deployment (MWAA/Composer/Helm)
■ CI/CD pipeline for DAGs
■ Monitoring & Slack alerting setup

AFTER THIS COURSE YOU CAN APPLY FOR:

→ Data Engineer (Airflow focus)
→ Analytics Engineer (dbt + Airflow)
→ Cloud Data Engineer (MWAA/Composer)
→ Senior Data Engineer / Lead
→ Data Platform Engineer
→ MLOps Engineer (ML pipeline orchestration)

COURSE SUMMARY & OUTCOMES

WHAT YOU'LL MASTER

✓ Design and deploy production-grade Airflow DAGs
✓ Integrate Airflow with AWS, Azure, and GCP services
✓ Orchestrate Spark, Snowflake, and dbt pipelines
✓ Master dynamic task mapping and DAG factory patterns
✓ Deploy Airflow on Kubernetes, MWAA, Composer, and Astronomer
✓ Implement CI/CD pipelines for DAG deployment
✓ Set up monitoring, alerting, and observability
✓ Apply security best practices — secrets, RBAC, VPC
✓ Build 5 end-to-end real-world projects
✓ Pass the Astronomer Certification for Apache Airflow

COURSE STATISTICS

• Total Duration: 110 hours (14-16 weeks)
• 21 Core Modules + 10 Bonus Topics
• 30+ Hands-On Labs
• 5 End-to-End Real-World Projects
• 200+ DAG examples & operator patterns
• Mock interviews & certification prep
• Lifetime Access to Course Materials
• Job Placement Assistance

ENROLLMENT & CONTACT

Website: databrickstraining.in
Email: info@databrickstraining.in
Phone: +91-8500002025

Training & Placement Excellence | *Trusted by 5000+ Data Engineers*