



Multi-Cloud Data Engineering Foundations

Azure (primary) • AWS • GCP

**Build, model, orchestrate and test modern data pipelines -
Azure-first, mapped to AWS and GCP**

50 Class Hours

45 Lab Hours

25 Sessions

5 Weeks

Class 18:00-20:00 IST, Monday to Friday • Labs self-paced • Instructor-led, live-online



Multi-Cloud Data Engineering Foundations

Azure (primary) • AWS • GCP

Instructor Information

Instructor	Email	Duration
ExaGuru	support@exaguru.com	50 class hours + 45 lab hours (5 weeks)

Delivery Model

Class timing	18:00 - 20:00 IST, Monday to Friday (2 contact hours per session)
Class hours per week	10 (5 sessions × 2 hours)
Lab hours per week	8-10, varies by week - fully self-paced from written lab guides
Structure	4 teaching weeks + 1 capstone week
Total	50 class hours + 45 lab hours = 95 hours over 5 weeks (25 class sessions)
Cloud weighting	Azure ~55% • AWS ~20% • GCP ~20% • neutral ~5% - Azure is the teaching vehicle; AWS and GCP are taught as mapped comparisons
Batch / mode	30+ learners, live-online

Session Format (2 hours)

Block	Minutes	What happens
Concept	40	Cloud-neutral data engineering concept. Whiteboard, no tooling. This is where the durable learning happens
Azure implementation	35	The concept built in the Azure service — ADF, Databricks, Fabric, Synapse. Instructor-driven demonstration
AWS + GCP mapping	15	The same concept mapped to the AWS and GCP equivalents, with the differences that actually matter
Structured Q&A	20	Written questions collected during the session, answered by priority
Lab briefing	10	Walk the week's lab guide, name the validation artefact, flag cost traps

Entry Requirements

- Working SQL: joins, aggregations, subqueries. Window functions are taught, not assumed.
- Basic Python: functions, loops, dicts, reading a file. PySpark is taught from zero.



- Comfortable on a command line and with Git basics.
- NO prior cloud experience assumed — Week 1 builds cloud storage from scratch.

No SQL or Python yet? Take a SQL + Python pre-course first - the Python and Spark sessions start in Week 1.

Who This Program Is For

Audience	Assumed Background	Why This Program Fits	Outcome Role
ETL / Informatica / SSIS developers	Strong SQL, mature ETL concepts	Best fit in the room. They know modelling and ELT already — this is a platform transition	Cloud Data Engineer (Azure-primary)
Database / BI developers	SQL, stored procedures, reporting	Strong fit. Needs to build Python and Spark from Week 1	Data Engineer, Analytics Engineer
Python developers / analysts	Python, some SQL	Good fit. Modelling and warehousing will be the new material	Data Engineer
Data analysts moving up	SQL, Power BI, business context	Viable. The Python and Spark weeks will be the hardest stretch	Analytics Engineer → Data Engineer
Cloud / DevOps engineers	Cloud, IaC, CI/CD	Good fit for the platform and DataOps side; must build SQL depth	Data Platform Engineer
Freshers with strong SQL + Python	Coursework, no production experience	Viable, but the full 18-20 hrs/week is genuinely required	Junior Data Engineer

Screening Test Before Enrolment (45 minutes)

Area	Sample Question Type	Pass Bar
SQL — joins	Given three tables, return customers with no orders in the last 90 days	Correct, using the right join type
SQL — aggregation	Return the top 3 products by revenue within each category	Correct; window function not required but a strong signal
SQL — reasoning	Why might this query return duplicate rows?	Identifies fan-out from a one-to-many join
Python	Read a CSV, filter rows, group and write out — without pandas	Working code using dicts and file handling
Python	What does this list comprehension produce?	Correct trace
Concepts	Explain the difference between a fact and a dimension	Coherent explanation, even if informal
Command line / Git	Clone a repo, create a branch, commit, push	Completes without help



Practice Environments (your own free-tier / trial accounts)

Platform	What You Get	Duration
Databricks Free Edition	Perpetually free, serverless-only. Small all-purpose compute, one 2X-Small SQL warehouse, max 5 concurrent job tasks, one active pipeline per type, one workspace and one metastore. Quota breach shuts compute off for the rest of the day. Non-commercial use only.	Perpetual
Microsoft Fabric trial	60-day trial capacity, 64 CU, up to 1 TB OneLake, full feature access except Copilot, private link and trusted workspaces. TENANT-SCOPED with a limit of 5 trial capacities per tenant.	60 days
Azure (subscription)	\$200 credit for the first 30 days, plus 12 months of selected free services after upgrading to pay-as-you-go.	Credits: 30 days
GCP	\$300 credit for 90 days, plus BigQuery always-free: roughly 1 TB of queries and 10 GB storage per month.	90 days + always-free
AWS	Credit-based since July 2025: \$100 on signup, up to \$100 more via onboarding. Some expensive services restricted on the Free Plan.	6 months or credits



Table of Contents

Program Overview & Course Details

Phase 1: Foundations (Weeks 1–4)

S01: The Data Engineering Landscape
S02: SQL for Data Engineers
S03: Python & the PySpark Mental Model
S04: File Formats & Storage Layout
S05: Cloud Object Storage
S06: Azure Data Factory Fundamentals
S07: ADF Ingestion Patterns
S08: Relational Sources & Change Data Capture
S09: AWS & GCP Ingestion Services
S10: Ingestion Design Clinic
S11: Spark Architecture
S12: PySpark Transformations
S13: Delta Lake
S14: Delta Optimisation & Table Maintenance
S15: Medallion Architecture Build
S16: Dimensional Modelling
S17: Azure Warehousing — Synapse & Fabric
S18: AWS Redshift & GCP BigQuery
S19: Orchestration & Scheduling
S20: Data Quality & Testing

Capstone Week (Week 5)

S21: Infrastructure as Code for Data Platforms
S22: Capstone Briefing & Pipeline Design Workshop
S23: Capstone Build Clinic
S24: Pipeline Debugging & Optimisation Drill
S25: Assessment & Design Defence
Mini-Projects & Capstones
Data Engineering Service Matrix (Azure ↔ AWS ↔ GCP)
Certification Alignment (optional)
Course Schedule (Module-wise Summary)

Expected Learning Outcomes

- Build batch pipelines on Azure: ADF ingestion and CDC, PySpark transformations and a Delta Lake medallion lakehouse
- Model data dimensionally and serve it through Synapse, Microsoft Fabric, Redshift and BigQuery
- Orchestrate, test and codify pipelines with retries, quality gates and Terraform, and debug them under pressure
- Map every concept to its AWS and GCP equivalent
- Deliver a capstone retail data platform and defend its design - ready for the Data Engineer (Azure-primary) role



Phase 1: Foundations (Weeks 1–4)

S01: The Data Engineering Landscape

2 Hours • Week 1, Mon

Topics Covered

- OLTP vs OLAP
- batch vs streaming
- ETL vs ELT and why ELT won
- lake vs warehouse vs lakehouse
- medallion architecture
- where the DE role stops and analyst/scientist begins

Cloud Implementation

- Azure (primary): Fabric / Databricks / Synapse positioning
- AWS: Redshift / EMR / Glue positioning
- GCP: BigQuery / Dataflow positioning
- Cloud focus: Cloud-neutral

Self-Paced Lab / Mini-Project

HANDS-ON LAB | Design • 1 hr self-paced

Map a real reporting requirement to an architecture on paper



S02: SQL for Data Engineers

2 Hours • Week 1, Tue

Topics Covered

- Beyond joins: window functions, CTEs, grouping sets, QUALIFY
- set operations
- reading a query plan
- why your query is slow
- idempotent SQL

Cloud Implementation

- Azure (primary): Azure SQL, Synapse serverless SQL
- AWS: Athena
- GCP: BigQuery
- Cloud focus: Cross-cloud

Self-Paced Lab / Mini-Project

HANDS-ON LAB | Hands-on • 2 hrs self-paced

Solve 15 window-function and plan-reading problems on a real dataset

S03: Python & the PySpark Mental Model

2 Hours • Week 1, Wed

Topics Covered

- Python for pipelines, not scripts
- pandas vs Spark and when NOT to use Spark
- lazy evaluation
- notebook discipline and why notebooks are not production

Cloud Implementation

- Azure (primary): Azure Databricks notebooks
- AWS: EMR / Glue notebooks
- GCP: Dataproc / Colab Enterprise
- Cloud focus: Cloud-neutral

Self-Paced Lab / Mini-Project

HANDS-ON LAB | Hands-on • 2 hrs self-paced

Rewrite a pandas script as PySpark; measure the crossover point



S04: File Formats & Storage Layout

2 Hours • Week 1, Thu

Topics Covered

- CSV / JSON / Parquet / Avro / ORC
- columnar vs row
- compression codecs
- predicate pushdown
- partitioning strategy
- the small-file problem
- target file sizing

Cloud Implementation

- Azure (primary): ADLS Gen2 layout conventions
- AWS: S3 layout conventions
- GCP: GCS layout conventions
- Cloud focus: Cross-cloud

Self-Paced Lab / Mini-Project

HANDS-ON LAB | Hands-on • 1 hr self-paced

Convert a CSV dataset to partitioned Parquet; measure scan cost and size

S05: Cloud Object Storage

2 Hours • Week 1, Fri

Topics Covered

- Object storage as the platform foundation
- hierarchical namespace and why it matters for analytics
- access models
- storage tiers and lifecycle
- cost drivers

Cloud Implementation

- Azure (primary): ADLS Gen2 — HNS, RBAC vs POSIX ACLs, tiers
- AWS: S3 — prefixes, storage classes, bucket policies
- GCP: GCS — buckets, classes, IAM
- Cloud focus: Azure-led

Self-Paced Lab / Mini-Project

HANDS-ON LAB | Hands-on • 2 hrs self-paced

Set up all three accounts WITH budget alerts; build an identical medallion folder layout



S06: Azure Data Factory Fundamentals

2 Hours • Week 2, Mon

Topics Covered

- Pipeline / activity / linked service / dataset model
- integration runtimes and when you need self-hosted
- triggers
- debug vs published runs

Cloud Implementation

- Azure (primary): ADF — full session
- AWS: Glue Studio / Step Functions (mapped)
- GCP: Data Fusion (mapped)
- Cloud focus: Azure-led

Self-Paced Lab / Mini-Project

HANDS-ON LAB | Hands-on • 2 hrs self-paced

Build a first ADF pipeline copying from HTTP source to ADLS bronze

S07: ADF Ingestion Patterns

2 Hours • Week 2, Tue

Topics Covered

- Copy activity at scale
- parameterisation
- incremental load and watermarking
- metadata-driven ingestion frameworks
- handling schema drift

Cloud Implementation

- Azure (primary): ADF — parameterised metadata-driven framework
- AWS: Glue job bookmarks (mapped)
- GCP: Data Fusion / Datastream (mapped)
- Cloud focus: Azure-led

Self-Paced Lab / Mini-Project

HANDS-ON LAB | MP-A • Hands-on • 2 hrs self-paced

Metadata-driven incremental ingestion framework in ADF



S08: Relational Sources & Change Data Capture

2 Hours • Week 2, Wed

Topics Covered

- Extraction patterns from OLTP
- full vs incremental vs CDC
- log-based CDC concepts
- handling deletes
- minimising source impact

Cloud Implementation

- Azure (primary): Azure SQL CDC, ADF change data capture
- AWS: DMS, Aurora CDC
- GCP: Datastream, Cloud SQL
- Cloud focus: Azure-led

Self-Paced Lab / Mini-Project

HANDS-ON LAB | Hands-on • 2 hrs self-paced

Configure CDC on a source table; land change feed into bronze

S09: AWS & GCP Ingestion Services

2 Hours • Week 2, Thu

Topics Covered

- The AWS and GCP ingestion stacks in depth
- service selection logic
- what each does better than ADF and where it is worse

Cloud Implementation

- Azure (primary): (comparison baseline)
- AWS: Glue, DMS, AppFlow, Transfer Family
- GCP: Data Fusion, Datastream, Storage Transfer, Dataflow templates
- Cloud focus: AWS + GCP

Self-Paced Lab / Mini-Project

HANDS-ON LAB | Hands-on • 2 hrs self-paced

Rebuild the S06 ingestion pipeline in Glue and in Data Fusion; compare effort



S10: Ingestion Design Clinic

2 Hours • Week 2, Fri

Topics Covered

- Choosing between file, API, database and stream ingestion
- idempotency and replay
- late-arriving and out-of-order data
- schema drift policy
- dead-letter design

Cloud Implementation

- Azure (primary): ADF patterns
- AWS: Glue patterns
- GCP: Data Fusion patterns
- Cloud focus: Cross-cloud

Self-Paced Lab / Mini-Project

HANDS-ON LAB | Design • 1 hr self-paced

Design ingestion for three written enterprise briefs and defend the choices

S11: Spark Architecture

2 Hours • Week 3, Mon

Topics Covered

- Driver and executors
- jobs, stages, tasks
- the shuffle and why it dominates runtime
- partitions
- Catalyst and adaptive query execution
- reading the Spark UI

Cloud Implementation

- Azure (primary): Azure Databricks / Fabric Spark
- AWS: EMR, Glue Spark
- GCP: Dataproc, Dataflow
- Cloud focus: Cloud-neutral

Self-Paced Lab / Mini-Project

HANDS-ON LAB | Hands-on • 2 hrs self-paced

Read the Spark UI for a deliberately bad query; identify the shuffle



S12: PySpark Transformations

2 Hours • Week 3, Tue

Topics Covered

- DataFrame API in depth
- joins and join strategies
- aggregations
- built-ins vs UDFs and the performance cliff
- handling nulls and bad records

Cloud Implementation

- Azure (primary): Databricks Free Edition notebooks
- AWS: EMR / Glue (mapped)
- GCP: Dataproc (mapped)
- Cloud focus: Azure-led

Self-Paced Lab / Mini-Project

HANDS-ON LAB | MP-B • Hands-on • 2 hrs self-paced

Transform raw bronze into cleaned, conformed silver tables

S13: Delta Lake

2 Hours • Week 3, Wed

Topics Covered

- The transaction log and how ACID actually works on object storage
- time travel
- MERGE for upserts
- schema enforcement vs evolution
- concurrency and conflict

Cloud Implementation

- Azure (primary): Delta on Databricks + Fabric
- AWS: Delta / Iceberg on EMR, Glue Iceberg tables
- GCP: Delta / Iceberg on Dataproc, BigLake
- Cloud focus: Azure-led

Self-Paced Lab / Mini-Project

HANDS-ON LAB | Hands-on • 2 hrs self-paced

Implement SCD Type 2 with MERGE; use time travel to recover a bad load



S14: Delta Optimisation & Table Maintenance

2 Hours • Week 3, Thu

Topics Covered

- OPTIMIZE and compaction
- Z-ORDER and liquid clustering
- VACUUM and retention
- partition strategy and over-partitioning
- statistics and data skipping

Cloud Implementation

- Azure (primary): Databricks OPTIMIZE, liquid clustering
- AWS: Iceberg compaction, Glue optimisation
- GCP: BigLake / Iceberg maintenance
- Cloud focus: Azure-led

Self-Paced Lab / Mini-Project

HANDS-ON LAB | Hands-on • 2 hrs self-paced

Measure query time before and after OPTIMIZE and clustering on the same table

S15: Medallion Architecture Build

2 Hours • Week 3, Fri

Topics Covered

- Bronze / silver / gold responsibilities
- idempotent reprocessing
- quarantine and dead-letter tables
- full vs incremental refresh
- reprocessing without downtime

Cloud Implementation

- Azure (primary): Databricks + ADLS medallion
- AWS: S3 + Glue medallion
- GCP: GCS + BigQuery medallion
- Cloud focus: Azure-led

Self-Paced Lab / Mini-Project

HANDS-ON LAB | MP-B • Hands-on • 2 hrs self-paced

Complete bronze→silver→gold pipeline, reprocessible end to end



S16: Dimensional Modelling

2 Hours • Week 4, Mon

Topics Covered

- Star schema
- declaring grain
- fact table types (transaction, periodic, accumulating)
- SCD Types 1/2/3
- surrogate keys
- conformed dimensions
- junk and degenerate dimensions

Cloud Implementation

- Azure (primary): Modelling for Synapse / Fabric
- AWS: Modelling for Redshift
- GCP: Modelling for BigQuery
- Cloud focus: Cloud-neutral

Self-Paced Lab / Mini-Project

HANDS-ON LAB | Design • 2 hrs self-paced

Model a retail scenario into a star schema and justify the grain

S17: Azure Warehousing — Synapse & Fabric

2 Hours • Week 4, Tue

Topics Covered

- Synapse dedicated vs serverless SQL pool
- Microsoft Fabric: OneLake, Lakehouse vs Warehouse, shortcuts, Direct Lake
- capacity units
- when to pick which

Cloud Implementation

- Azure (primary): Synapse + Microsoft Fabric — full session
- AWS: (mapped in S18)
- GCP: (mapped in S18)
- Cloud focus: Azure-led

Self-Paced Lab / Mini-Project

HANDS-ON LAB | MP-C • Hands-on • 2 hrs self-paced

Serve the gold layer through a Fabric Lakehouse and a Warehouse; compare



S18: AWS Redshift & GCP BigQuery

2 Hours • Week 4, Wed

Topics Covered

- Redshift architecture, distribution and sort keys, RA3, Serverless
- BigQuery architecture, slots, partitioning and clustering
- cost models compared
- when each genuinely wins

Cloud Implementation

- Azure (primary): (comparison baseline)
- AWS: Redshift, Redshift Serverless
- GCP: BigQuery, BI Engine
- Cloud focus: AWS + GCP

Self-Paced Lab / Mini-Project

HANDS-ON LAB | Mixed • 2 hrs self-paced

Load the same gold dataset into BigQuery and Redshift Serverless; compare cost and speed

S19: Orchestration & Scheduling

2 Hours • Week 4, Thu

Topics Covered

- DAGs and dependency management
- triggers vs schedules
- retries, backfills and catch-up
- SLAs and alerting
- idempotent task design
- parameter passing

Cloud Implementation

- Azure (primary): ADF triggers, Fabric Data Pipelines
- AWS: Airflow / MWAA, Step Functions
- GCP: Cloud Composer, Workflows
- Cloud focus: Azure-led

Self-Paced Lab / Mini-Project

HANDS-ON LAB | Hands-on • 2 hrs self-paced

Orchestrate the full medallion pipeline with dependencies, retries and failure alerts



S20: Data Quality & Testing

2 Hours • Week 4, Fri

Topics Covered

- What to test in a pipeline
- null, uniqueness, referential and range checks
- reconciliation against source
- expectations frameworks
- failing loud vs quarantining

Cloud Implementation

- Azure (primary): Great Expectations / dbt tests on Databricks
- AWS: Deequ, Glue Data Quality
- GCP: Dataplex data quality
- Cloud focus: Cross-cloud

Self-Paced Lab / Mini-Project

HANDS-ON LAB | MP-C • Hands-on • 2 hrs self-paced

Add a quality gate that blocks bad data from reaching gold

Capstone Week (Week 5)

S21: Infrastructure as Code for Data Platforms

2 Hours • Week 5, Mon

Topics Covered

- Why data infrastructure belongs in code
- provider model
- what to codify and what not to
- environment promotion
- secrets handling

Cloud Implementation

- Azure (primary): Terraform / Bicep for ADLS, ADF, Databricks
- AWS: Terraform for S3, Glue
- GCP: Terraform for GCS, BigQuery
- Cloud focus: Azure-led

Self-Paced Lab / Mini-Project

HANDS-ON LAB | Hands-on • 2 hrs self-paced

Codify the storage and warehouse layer in Terraform



S22: Capstone Briefing & Pipeline Design Workshop

2 Hours • Week 5, Tue

Topics Covered

- Scenario: retail company, three source systems, daily and hourly SLAs
- requirements to architecture
- modelling decisions
- peer design review

Cloud Implementation

- Azure (primary): Azure design track
- AWS: AWS mapping
- GCP: GCP mapping
- Cloud focus: Cross-cloud

Self-Paced Lab / Mini-Project

HANDS-ON LAB | CAP-1 • Design • 1 hr self-paced
architecture document, data model and SLA plan

S23: Capstone Build Clinic

2 Hours • Week 5, Wed

Topics Covered

- Instructor-led build session
- unblocking
- code review of learner pipelines
- validation strategy

Cloud Implementation

- Azure (primary): ADF + Databricks + Fabric
- AWS: (optional AWS track)
- GCP: (optional GCP track)
- Cloud focus: Azure-led

Self-Paced Lab / Mini-Project

HANDS-ON LAB | CAP-1 • Hands-on • 2 hrs self-paced
guided build; bulk of construction happens in self-paced lab hours



S24: Pipeline Debugging & Optimisation Drill

2 Hours • Week 5, Thu

Topics Covered

- Instructor injects failures: skew, schema drift, duplicate loads, late data, a runaway cluster
- learners diagnose under time pressure

Cloud Implementation

- Azure (primary): Spark UI, ADF monitoring
- AWS: CloudWatch, Glue metrics
- GCP: Dataproc / BigQuery job stats
- Cloud focus: Azure-led

Self-Paced Lab / Mini-Project

HANDS-ON LAB | Hands-on • 2 hrs self-paced

Timed fault diagnosis; produce a production runbook

S25: Assessment & Design Defence

2 Hours • Week 5, Fri

Topics Covered

- Written knowledge test
- hands-on validation of the capstone
- each learner defends their pipeline design and its trade-offs

Cloud Implementation

- Cloud focus: Cross-cloud

Self-Paced Lab / Mini-Project

HANDS-ON LAB | Mixed • 1 hr self-paced

Assessment + 10-minute design defence per learner

Mini-Projects & Capstones

Four mini-projects plus a capstone, all fully self-paced from written lab guides.

Week	Code	Mini-Project / Capstone	What You Produce	Mode	Lab Hrs
Wk 1	—	Foundations warm-up	Window-function problem set, a pandas-to-PySpark rewrite with a measured crossover point, partitioned Parquet conversion, and three cloud accounts with budget alerts live	Hands-on	8
Wk 2	MP-A	Metadata-driven ingestion	A parameterised ADF framework that ingests multiple sources incrementally with watermarking, plus the same pipeline rebuilt in Glue and Data Fusion for comparison	Hands-on	9



Wk 3	MP-B	Medallion lakehouse	Bronze→silver→gold on Delta with SCD Type 2 via MERGE, idempotent reprocessing, quarantine tables, and a measured before/after OPTIMIZE comparison	Hands-on	10
Wk 4	MP-C	Modelled warehouse + quality gate	Star schema served through a Fabric Lakehouse and Warehouse, the same gold loaded into BigQuery, full orchestration with retries, and a quality gate that blocks bad data	Hands-on	10
Wk 5	CAP-1	CAPSTONE — Retail Data Platform	Three source systems to a governed star schema: ingestion framework, medallion lakehouse, dimensional model, orchestration with SLAs, quality gates, Terraform, runbook and architecture diagram	Hands-on	8
					45

How the self-paced labs work

- Every mini-project ships with a written lab guide: prerequisites, numbered steps per platform, expected output at each step, a validation checklist, a teardown step and a troubleshooting appendix.
- Every mini-project has an objective validation artefact the learner submits — a notebook export, a table row count, a dbt test result, a Spark UI screenshot. Not 'I finished it'.
- Learners are assigned to pods of 5-6 who peer-review weekly against a published rubric. This is the only thing that scales review to 30+ learners.
- Datasets are provided, not sourced by learners. Use one consistent dataset across the whole program so learners build depth instead of restarting.

Data Engineering Service Matrix (Azure ↔ AWS ↔ GCP)

Quick-reference card: the Azure column comes first because Azure is the teaching vehicle.

Capability	Azure (primary)	AWS	GCP	Design Note
Object storage / data lake	ADLS Gen2	S3	Cloud Storage	Only ADLS has a true hierarchical namespace — it changes directory operations and ACLs
Batch ingestion / ETL orchestration	Azure Data Factory	Glue, Step Functions	Data Fusion, Workflows	ADF is the most GUI-driven; Glue is the most code-first
Database migration & CDC	ADF CDC, Azure SQL CDC	DMS	Datastream	Log-based CDC is the only pattern that scales without hurting the source
Spark compute	Azure Databricks, Fabric Spark	EMR, Glue Spark	Dataproc	Databricks is the reference implementation and the free-tier path
Open table format	Delta Lake	Iceberg (Glue), Delta	Iceberg / Delta via BigLake	Delta on Azure, Iceberg increasingly the AWS default — know both transaction logs



Serverless SQL over the lake	Synapse serverless SQL	Athena	BigQuery external tables	Pay-per-TB-scanned everywhere; partitioning is what controls your bill
Cloud data warehouse	Synapse dedicated, Fabric Warehouse	Redshift, Redshift Serverless	BigQuery	BigQuery separates storage and compute most cleanly; Redshift needs dist/sort key design
Lakehouse platform	Microsoft Fabric, Databricks	Databricks on AWS, Lake Formation	Databricks on GCP, BigLake	Fabric's OneLake + Direct Lake is genuinely differentiated — no true AWS or GCP equivalent
Workflow orchestration	ADF triggers, Fabric Data Pipelines	MWAA (Airflow), Step Functions	Cloud Composer (Airflow)	Airflow is the portable skill; the managed versions are expensive to run for learning
Data quality	Great Expectations, dbt tests	Glue Data Quality, Deequ	Dataplex data quality	Framework choice matters less than deciding what to test and when to fail

Certification Alignment (optional)

Azure (primary)	AWS	GCP	Notes
DP-203 successor / DP-700 (Fabric Data Engineer)	Data Engineer – Associate	Associate Cloud Engineer	Not exam prep; roughly 60-70% overlap with DP-700
Databricks Certified Data Engineer Associate	—	—	Strongly recommended. Databricks Free Edition covers most of the practical preparation

----- End of Syllabus -----





Course Schedule

Module	Topic
Phase 1: Foundations (Weeks 1–4)	
S01 The Data Engineering Landscape 2 Hours • Week 1, Mon	The Data Engineering Landscape • OLTP vs OLAP • batch vs streaming • ETL vs ELT and why ELT won • lake vs warehouse vs lakehouse • Hands-on: Map a real reporting requirement to an architecture on paper
S02 SQL for Data Engineers 2 Hours • Week 1, Tue	SQL for Data Engineers • Beyond joins: window functions, CTEs, grouping sets, QUALIFY • set operations • reading a query plan • why your query is slow • Hands-on: Solve 15 window-function and plan-reading problems on a real dataset
S03 Python & the PySpark Mental Model 2 Hours • Week 1, Wed	Python & the PySpark Mental Model • Python for pipelines, not scripts • pandas vs Spark and when NOT to use Spark • lazy evaluation • notebook discipline and why notebooks are not production • Hands-on: Rewrite a pandas script as PySpark; measure the crossover point
S04 File Formats & Storage Layout 2 Hours • Week 1, Thu	File Formats & Storage Layout • CSV / JSON / Parquet / Avro / ORC • columnar vs row • compression codecs • predicate pushdown • Hands-on: Convert a CSV dataset to partitioned Parquet; measure scan cost and size



S05

Cloud Object Storage

2 Hours • Week 1, Fri

Cloud Object Storage

- Object storage as the platform foundation
- hierarchical namespace and why it matters for analytics
- access models
- storage tiers and lifecycle
- Hands-on: Set up all three accounts WITH budget alerts; build an identical medallion folder layout

S06

Azure Data Factory Fundamentals

2 Hours • Week 2, Mon

Azure Data Factory Fundamentals

- Pipeline / activity / linked service / dataset model
- integration runtimes and when you need self-hosted
- triggers
- debug vs published runs
- Hands-on: Build a first ADF pipeline copying from HTTP source to ADLS bronze

S07

ADF Ingestion Patterns

2 Hours • Week 2, Tue

ADF Ingestion Patterns

- Copy activity at scale
- parameterisation
- incremental load and watermarking
- metadata-driven ingestion frameworks
- Hands-on: Metadata-driven incremental ingestion framework in ADF

S08

Relational Sources & Change Data Capture

2 Hours • Week 2, Wed

Relational Sources & Change Data Capture

- Extraction patterns from OLTP
- full vs incremental vs CDC
- log-based CDC concepts
- handling deletes
- Hands-on: Configure CDC on a source table; land change feed into bronze

S09

AWS & GCP Ingestion Services

2 Hours • Week 2, Thu

AWS & GCP Ingestion Services

- The AWS and GCP ingestion stacks in depth
- service selection logic
- what each does better than ADF and where it is worse
- Azure (primary): (comparison baseline)
- Hands-on: Rebuild the S06 ingestion pipeline in Glue and in Data Fusion; compare effort



S10

Ingestion Design Clinic

2 Hours • Week 2, Fri

Ingestion Design Clinic

- Choosing between file, API, database and stream ingestion
- idempotency and replay
- late-arriving and out-of-order data
- schema drift policy
- Hands-on: Design ingestion for three written enterprise briefs and defend the choices

S11

Spark Architecture

2 Hours • Week 3, Mon

Spark Architecture

- Driver and executors
- jobs, stages, tasks
- the shuffle and why it dominates runtime
- partitions
- Hands-on: Read the Spark UI for a deliberately bad query; identify the shuffle

S12

PySpark Transformations

2 Hours • Week 3, Tue

PySpark Transformations

- DataFrame API in depth
- joins and join strategies
- aggregations
- built-ins vs UDFs and the performance cliff
- Hands-on: Transform raw bronze into cleaned, conformed silver tables

S13

Delta Lake

2 Hours • Week 3, Wed

Delta Lake

- The transaction log and how ACID actually works on object storage
- time travel
- MERGE for upserts
- schema enforcement vs evolution
- Hands-on: Implement SCD Type 2 with MERGE; use time travel to recover a bad load

S14

Delta Optimisation & Table Maintenance

2 Hours • Week 3, Thu

Delta Optimisation & Table Maintenance

- OPTIMIZE and compaction
- Z-ORDER and liquid clustering
- VACUUM and retention
- partition strategy and over-partitioning
- Hands-on: Measure query time before and after OPTIMIZE and clustering on the same table



S15

Medallion Architecture Build

2 Hours • Week 3, Fri

Medallion Architecture Build

- Bronze / silver / gold responsibilities
- idempotent reprocessing
- quarantine and dead-letter tables
- full vs incremental refresh
- Hands-on: Complete bronze→silver→gold pipeline, reprocessible end to end

S16

Dimensional Modelling

2 Hours • Week 4, Mon

Dimensional Modelling

- Star schema
- declaring grain
- fact table types (transaction, periodic, accumulating)
- SCD Types 1/2/3
- Hands-on: Model a retail scenario into a star schema and justify the grain

S17

Azure Warehousing — Synapse & Fabric

2 Hours • Week 4, Tue

Azure Warehousing — Synapse & Fabric

- Synapse dedicated vs serverless SQL pool
- Microsoft Fabric: OneLake, Lakehouse vs Warehouse, shortcuts, Direct Lake
- capacity units
- when to pick which
- Hands-on: Serve the gold layer through a Fabric Lakehouse and a Warehouse; compare

S18

AWS Redshift & GCP BigQuery

2 Hours • Week 4, Wed

AWS Redshift & GCP BigQuery

- Redshift architecture, distribution and sort keys, RA3, Serverless
- BigQuery architecture, slots, partitioning and clustering
- cost models compared
- when each genuinely wins
- Hands-on: Load the same gold dataset into BigQuery and Redshift Serverless; compare cost and speed

S19

Orchestration & Scheduling

2 Hours • Week 4, Thu

Orchestration & Scheduling

- DAGs and dependency management
- triggers vs schedules
- retries, backfills and catch-up
- SLAs and alerting
- Hands-on: Orchestrate the full medallion pipeline with dependencies, retries and failure alerts



S20 Data Quality & Testing 2 Hours • Week 4, Fri	Data Quality & Testing • What to test in a pipeline • null, uniqueness, referential and range checks • reconciliation against source • expectations frameworks • Hands-on: Add a quality gate that blocks bad data from reaching gold
Capstone Week (Week 5)	
S21 Infrastructure as Code for Data Platforms 2 Hours • Week 5, Mon	Infrastructure as Code for Data Platforms • Why data infrastructure belongs in code • provider model • what to codify and what not to • environment promotion • Hands-on: Codify the storage and warehouse layer in Terraform
S22 Capstone Briefing & Pipeline Design Workshop 2 Hours • Week 5, Tue	Capstone Briefing & Pipeline Design Workshop • Scenario: retail company, three source systems, daily and hourly SLAs • requirements to architecture • modelling decisions • peer design review • Hands-on: architecture document, data model and SLA plan
S23 Capstone Build Clinic 2 Hours • Week 5, Wed	Capstone Build Clinic • Instructor-led build session • unblocking • code review of learner pipelines • validation strategy • Hands-on: guided build; bulk of construction happens in self-paced lab hours
S24 Pipeline Debugging & Optimisation Drill 2 Hours • Week 5, Thu	Pipeline Debugging & Optimisation Drill • Instructor injects failures: skew, schema drift, duplicate loads, late data, a runaway cluster • learners diagnose under time pressure • Azure (primary): Spark UI, ADF monitoring • AWS: CloudWatch, Glue metrics • Hands-on: Timed fault diagnosis; produce a production runbook



S25

Assessment & Design Defence

2 Hours • Week 5, Fri

Assessment & Design Defence

- Written knowledge test
- hands-on validation of the capstone
- each learner defends their pipeline design and its trade-offs
- Cloud focus: Cross-cloud
- Hands-on: Assessment + 10-minute design defence per learner

TOTAL

50 Class Hours + 45 Lab Hours



What Our Learners Say

Honest feedback from professionals and learners who have been part of the ExaGuru learning journey.



Excellent Online training institute for Exadata, Exacc and OCI

Excelent Online training institute for Exadata, Exacc and oci. I had a great learning experience. Special...

SK Sher Khan
UAE · Oct 9, 2026



It's very good platform to upgrade your skills

It's very good platform to upgrade your skills in very less time with hands-on practical knowledge as theory is not...

MS Mayuresh Shingan
India · Oct 9, 2026



Excellent Course with Real-Time Troubleshooting

My experience with the ExaGuru course has been very positive and valuable. The training covers Exadata, OCI, ExaCC...

SS Shaik Sulaiman
Singapore · Oct 2, 2026



Quite literally the most in dept educational content available...

Well I scoured the internet for in depth ExaData courses and couldn't find any then I stumbled upon Exaguru was a bit...

AS Astitva Singh
India · Aug 20, 2026



The finest Exadata training I found

I found Exaguru while checking all the other options available for Exadata training, and found their training program...

AY Aman Yadav
India · May 28, 2026



I was looking for OCI GUI access

I was looking for OCI GUI access. I got it here. I was able to explore the options in OCI. The explanation during the...

MM Moizuddin Mudassar
Saudi Arabia · Oct 2, 2026



I joined it in july batch

I joined it in july batch, I liked the OCI course, especially the networking labs. Working with VCNs, subnets and...

GT Gaurav Tiwari
India · Aug 20, 2026



Very good theory and practice sessions

Very good theory and practice sessions. The instructor is very knowledgeable and took sincere efforts to make the...

AA Amit Aggarwal
United States · May 27, 2026



Here I made progress in my OCI Journey

Here I made progress in my OCI Journey, all thanks mainly goes to Krishna Sir, He guided me through and made me what...

AT Anuj Tiwari
India · Aug 22, 2026



Krishna is very good

Krishna is very good, patience enough to give answers give complete knowledge.

S Srinivas
United States · Oct 2, 2026



Nice OCI networking and Exadata class

Nice oci networking part and overall exadata class is good.

MK Muthu Kumar
India · Oct 2, 2026



Simple learning process

Simple learning process. Best tutor. Company should more focus on helping candidate get the job who participated in...

RY Rama Yeligeti
India · Jul 10, 2026



Amazing stuff and support provided by the team

Amazing stuff and support provided by team, recommended if want to reach pro level.

SC Shailendra Chauhan
India · Jul 7, 2026



It was an awesome class of Exadata and OCI

It was awesome class of exadata and OCI. Please keep good work continuesly.

HR Heeralal Ram
India · Jun 16, 2026

[Read more reviews on Trustpilot ↗](#)

<https://www.trustpilot.com/review/exaguru.com>





Take the Next Step in Your Career

Visit our Job Portal to explore openings, build your profile and connect with employers who are hiring Oracle Cloud and Exadata professionals.

Visit our Job Portal → elevate.exaguru.com

support@exaguru.com · www.exaguru.com

