



PrepNPlaced

LEARN · BUILD · GET PLACED

COHORT 04 · PROGRAM SYLLABUS

LIVE WEEKEND COHORT · ENGINEERING ANALYTICS

AI-Powered Data Engineering

ENGINEERING ANALYTICS · COHORT 04

The complete Data Engineering stack — SQL to Spark to Snowflake to Kafka to AWS — taught the 2026 way: with AI woven into every module, now with a dedicated interview & systems-readiness module, ending in Agentic AI.

192 hrs

LIVE LECTURES

8

MODULES · AI IN EVERY ONE

32 topics

6 HRS EACH · 2 CLASSES

+9 hrs

3 PROJECTS · ON GITHUB

₹59,000

COHORT 04 PRICE

40 seats

PER COHORT

⚡ AI-first: every module carries an “AI in Practice” thread — the syllabus builds toward the Agentic AI capstone.

MODULE 08 = CAPSTONE

Instructors: **Durgesh Yadav** (Founder & Lead Instructor) · **Pragya Rathi** (Co-Founder & Lead Instructor)

LIVE WEEKEND COHORT · DOUBT SUPPORT · REPLAYS

01 SQL & Analytics Foundation

The querying, modeling, and BI base every data role is screened on.

24 HRS · 4 TOPICS

SQL for Analytics

6 HRS · 2 CLASSES

- > Query execution order & how the engine runs SQL
- > Window functions — ROW_NUMBER, RANK, LAG / LEAD
- > CTEs, recursive & correlated subqueries
- > INNER / LEFT / SELF / ANTI joins
- > Running totals, moving averages & partitions
- > Query optimization, indexing & execution plans

Data Modeling & Warehousing Concepts

6 HRS · 2 CLASSES

- > Star vs Snowflake schema
- > Slowly Changing Dimensions (SCD Type 1 & 2)
- > Grain, surrogate keys & cardinality
- > Fact & dimension tables
- > Normalization vs denormalization
- > Data Vault (hub-satellite-link) modeling

Python for Data Analytics

6 HRS · 2 CLASSES

- > Pandas wrangling, joins & group-by
- > Data cleaning & missing-value handling
- > Matplotlib & Seaborn visualization
- > NumPy vectorization
- > Exploratory Data Analysis (EDA) & profiling

Business Intelligence with Power BI

6 HRS · 2 CLASSES

- > Data modeling & relationships
- > Time intelligence — YoY, MTD, YTD
- > Dashboard storytelling & UX
- > DAX measures & calculated columns
- > Row-Level Security (RLS)

< AI IN PRACTICE

LLM-assisted query writing and explaining execution plans with AI — and learning to verify what the AI got wrong.

02 Big Data & Apache Spark

Process data at scale the way product companies actually run it.

24 HRS · 4 TOPICS

Distributed Computing Foundations

6 HRS · 2 CLASSES

- > Partitioning & parallelism
- > Shuffle, wide vs narrow transformations
- > Cluster architecture — driver & executors
- > Lazy evaluation & the DAG

Apache Spark Core

6 HRS · 2 CLASSES

- > RDD vs DataFrame vs Dataset
- > Catalyst optimizer & Tungsten engine
- > Transformations & actions
- > Spark SQL & schema handling

PySpark for Engineering

6 HRS · 2 CLASSES

- > DataFrame API & joins at scale
- > UDFs & pandas UDFs
- > Aggregations & window functions
- > Performance tuning — broadcast joins, caching, AQE

Databricks & Delta Lake

6 HRS · 2 CLASSES

- > Lakehouse architecture
- > OPTIMIZE & Z-ordering
- > Delta Lake vs Apache Iceberg — table formats compared
- > Delta Lake ACID & time travel
- > Unity Catalog & governance

< A I I N P R A C T I C E

Using AI to diagnose shuffle-heavy jobs and read Spark query plans faster.

03 Modern Data Engineering

Snowflake, dbt, and Airflow — the modern data stack employers ask for.

24 HRS · 4 TOPICS

Cloud Data Warehousing (Snowflake)

6 HRS · 2 CLASSES

- > Virtual warehouses & elastic scaling
- > Streams & Tasks for CDC
- > Cost & performance optimization
- > Micro-partitions & clustering
- > Zero-copy cloning & time travel

Analytics Engineering with dbt

6 HRS · 2 CLASSES

- > Models, sources & ref()
- > Tests, docs & data lineage
- > CI for dbt projects
- > Incremental & materialized models
- > Snapshots, macros & Jinja

Orchestration with Airflow

6 HRS · 2 CLASSES

- > DAG authoring & task dependencies
- > Scheduling, backfills & catchup
- > Managed Airflow — MWAA & Cloud Composer
- > Operators, sensors & hooks
- > Idempotency, retries & SLAs
- > Dagster & Prefect as modern alternatives

Pipeline & Data Quality Patterns

6 HRS · 2 CLASSES

- > Medallion — bronze / silver / gold
- > Change Data Capture (CDC)
- > ETL vs ELT
- > Data quality, validation & contracts

< AI IN PRACTICE

AI-generated dbt tests and documentation, reviewed like a pull request.

04 Real-Time Streaming

Kafka and Spark Structured Streaming — the in-demand real-time skillset.

24 HRS · 4 TOPICS

Streaming Foundations

6 HRS · 2 CLASSES

- > Batch vs stream processing
- > Windowing & watermarks
- > Event time vs processing time
- > Delivery semantics — at-least / exactly-once

Apache Kafka

6 HRS · 2 CLASSES

- > Topics, partitions & offsets
- > Replication & fault tolerance
- > Producers, consumers & consumer groups
- > Schema Registry & Avro

Spark Structured Streaming

6 HRS · 2 CLASSES

- > readStream / writeStream
- > Kafka ↔ Delta Lake integration
- > Stateful aggregations & stream joins
- > Checkpointing & failure recovery

Real-Time Analytics & Serving

6 HRS · 2 CLASSES

- > Streaming into the lakehouse
- > Latency vs throughput tradeoffs
- > Live dashboards & alerting
- > Monitoring & observability

◀ A I I N P R A C T I C E

AI-assisted incident debugging on streaming failures — checkpoint recovery, lag analysis.

05 Data Engineering on AWS

24 HRS · 4 TOPICS

Ship data workloads on the cloud platform most job posts require — built around 5 core services: S3, Glue, Redshift, Kinesis, Lambda.

Amazon S3 — The Data Lake Foundation

6 HRS · 2 CLASSES

- > Buckets, objects, prefixes & partition layouts
- > S3 as a lakehouse landing zone (Parquet)
- > Storage classes & lifecycle policies
- > Security — IAM, encryption & access points

AWS Glue — Serverless ETL & Catalog

6 HRS · 2 CLASSES

- > Glue Data Catalog & crawlers (schema discovery)
- > Job bookmarks, triggers & workflows
- > Glue ETL jobs with PySpark
- > Schema evolution & partitioning

Amazon Redshift — Cloud Warehouse

6 HRS · 2 CLASSES

- > Cluster architecture, RA3 & Serverless
- > Redshift Spectrum — querying S3 directly
- > Workload management & cost tuning
- > Distribution keys & sort keys
- > COPY / UNLOAD & loading patterns

Kinesis + Lambda — Streaming & Serverless

6 HRS · 2 CLASSES

- > Kinesis Data Streams & shards
- > AWS Lambda event-driven processing
- > Triggers, IAM roles & monitoring
- > Kinesis Firehose → S3 / Redshift delivery
- > Event pipelines — S3 event → Lambda → Glue

◀ A I I N P R A C T I C E

Infrastructure-as-prompt — drafting Glue jobs and IAM policies with AI, then hardening them by hand.

06 Hadoop, Hive & Azure Data Engineering

Legacy big data and the Azure stack — the platforms enterprise interviews still cover.

24 HRS · 4 TOPICS

Hadoop & Big Data Foundation

6 HRS · 2 CLASSES

- > HDFS — blocks, replication & NameNode
- > MapReduce & YARN
- > Hadoop vs the modern lakehouse
- > Where legacy stacks appear in interviews

Hive & SQL-on-Hadoop

6 HRS · 2 CLASSES

- > Hive metastore
- > Partitioning & bucketing
- > SerDes
- > ORC vs Parquet
- > Query tuning

Azure Data Engineering Core

6 HRS · 2 CLASSES

- > ADLS Gen2
- > Azure Data Factory — pipelines & integration runtimes
- > Synapse — dedicated vs serverless
- > Azure Databricks

Azure Streaming & Operations

6 HRS · 2 CLASSES

- > Event Hubs
- > Stream Analytics
- > Azure Monitor & cost management
- > Purview governance

◀ A I I N P R A C T I C E

Interview prep with AI — drilling the legacy-stack questions interviewers still ask.

07 Interview & Systems Readiness

NEW IN COHORT 4

24 HRS · 4 TOPICS

System design, coding rounds, NoSQL, and DevOps — the rounds that decide the offer.

Data Engineering System Design

6 HRS · 2 CLASSES

- > Designing pipelines for scale — ingestion, transformation, serving layers
- > Capacity & throughput estimation
- > Whiteboard-style design walkthroughs on real interview scenarios
- > Batch vs streaming architecture trade-offs
- > Handling backfills, replays & late-arriving data

DSA & Python for Data Engineering Interviews

6 HRS · 2 CLASSES

- > Python coding-round patterns (not generic CS-DSA) — arrays, hashmaps, generators
- > Writing production-grade pipeline code — error handling, edge cases
- > Memory-efficient data processing patterns
- > Debugging scenario questions (“your pipeline ran but produced wrong data — investigate”)

NoSQL & Alternative Data Stores

6 HRS · 2 CLASSES

- > Document stores (MongoDB) vs wide-column (Cassandra) vs key-value (Redis, DynamoDB)
- > Elasticsearch for log/search pipelines
- > When NOT to use a relational store
- > NoSQL in real DE pipelines — CDC targets, caching layers

DevOps & Governance for Data Pipelines

6 HRS · 2 CLASSES

- > Docker & containerizing pipelines / Airflow / Spark jobs
- > CI/CD for dbt models & pipeline code
- > Terraform basics for provisioning data infra
- > Data contracts, PII masking & compliance basics (GDPR-style)

◀ A I I N P R A C T I C E

AI-assisted system design critique — feeding your pipeline design to AI for a mock architecture review before the real interview.

08

AI Automation & Agentic AI

THE CAPSTONE

24 HRS · 4 TOPICS

Use LLMs and AI agents to automate real data work — the 2026 edge.

Prompt Engineering for Data

6 HRS · 2 CLASSES

- > Prompt patterns & roles
- > Few-shot & chain-of-thought
- > Structured / JSON outputs
- > Guardrails & evaluation

AI-Powered Analytics

6 HRS · 2 CLASSES

- > LLM-assisted SQL & text-to-SQL
- > Automated reporting & narratives
- > RAG over your data docs
- > Embeddings & vector search

Agentic AI Foundations

6 HRS · 2 CLASSES

- > Agentic vs generative AI
- > Reasoning & planning loops
- > Memory, tools & function calling
- > Multi-agent collaboration

Agent Frameworks in Practice

6 HRS · 2 CLASSES

- > LangChain, LangGraph & CrewAI
- > Building a data-analyst agent
- > Tool & API integration
- > Evaluation & safety

THE CAPSTONE

Every "AI in Practice" thread from Modules 1–7 converges here: you go from using AI inside data work to building the agents that do it.

9 HOURS · PUBLIC GITHUB PROOF

Three production-style projects you ship

Beyond the 192 lecture hours, you build three real projects (9 guided hours, 3 hrs each). Each ships to GitHub so recruiters can open the repo and verify your work — not just read a resume bullet.

StreamCommerce Lakehouse 360

3 HRS · GITHUB REPO

A real-time e-commerce lakehouse: stream orders and clickstream through a bronze → silver → gold medallion pipeline and serve live revenue analytics.

KAFKA

SPARK STRUCTURED STREAMING

DELTA LAKE

DATABRICKS

Retail Revenue Intelligence

3 HRS · GITHUB REPO

End-to-end retail analytics: model the data, compute revenue, cohort, and retention metrics, and ship an executive dashboard.

SQL

PYTHON

PANDAS

POWER BI

Snowflake + dbt Analytics Engineering

3 HRS · GITHUB REPO

A modern ELT warehouse: build tested, documented dbt models and incremental pipelines into dimensional marts on Snowflake.

SNOWFLAKE

DBT

SQL

ELT

INSTRUCTORS

Taught by working data engineers

Durgesh Yadav

FOUNDER & LEAD INSTRUCTOR

Senior Data Engineer at 7-Eleven · ex-Product Analyst at Target

Pragya Rath

CO-FOUNDER & LEAD INSTRUCTOR

Data Engineer at 7-Eleven

ENROLLMENT · COHORT 04

₹59,000 40 seats per cohort

Live weekend classes · doubt support · replays

192 HRS · 8 MODULES · 32 TOPICS

Who should enroll

- + Learners moving toward data analyst, analytics engineer, or data engineering roles
- + Freshers who need SQL, Power BI, Python, PySpark, and project proof in one sequence
- + Working professionals who want structured assignments, real projects, and interview practice

Who should not enroll

- Anyone looking for a passive certificate without assignments or project work
- Candidates expecting guaranteed placement, salary, or interview outcomes
- Learners who are not ready to revise, submit work, and practice explanations

CAREER OS · FROM PROOF TO PLACEMENT

Resume AI

Rewrite course projects into honest resume bullets after you build them.

ATS Resume Checker

Check whether the course proof matches a target job description.

AI Mock Interview

Practice explaining SQL, dashboards, PySpark, pipelines, and projects.

Open Learning

Use free lessons to strengthen foundations before or during the cohort.