Data Engineering

Build reliable batch and streaming data pipelines with Spark, Databricks and the Hadoop ecosystem.

DURATION	FORMAT	LEVEL	LABS
45 hours	Onsite, online or hybrid	Intermediate	Cloud sandbox included

About this programme

Covers what engineers need to collect, store, transform and serve data at scale: data lakes and warehouses, Databricks and Apache Spark, the Hadoop ecosystem and Kafka, with governance, security and cloud cost practices built in.

Who it is for: Software engineers, ETL developers and data analysts moving into data engineering

What your team will be able to do

- Design ingestion, storage and transformation layers for lakes and warehouses
- Build pipelines with Databricks, PySpark and Delta Lake
- Process streaming data with Spark Streaming and Kafka
- Apply data quality, lineage and access control
- Optimise pipeline performance and cloud costs

Before you start

- Working SQL
- Basic Python or another programming language

Course outline

1. Data engineering foundations

- Big data characteristics: volume, velocity, variety and veracity
- Batch and real-time analytics
- Ingestion, storage, transformation and ETL pipelines
- Data warehouses and data lakes
- Cloud platforms (AWS, Azure, GCP) for data engineering

2. Apache Spark and PySpark

- Spark architecture: RDDs, DataFrames and Datasets
- Spark SQL and Spark Streaming
- PySpark DataFrames and transformations
- Spark MLlib and GraphX overview

3. Databricks

- Databricks notebooks and clusters
- Databricks SQL and Delta Lake
- Jobs and workflows

4. The Hadoop ecosystem

- HDFS architecture and commands
- MapReduce programming with Python
- YARN resource management
- Data movement with Sqoop, Flume and Kafka

5. Governance, visualisation and best practice

- Data quality and data lineage
- Data security and access control
- Visualising data with Tableau or Power BI
- Optimising pipelines and managing cloud costs

6. Projects

- Building data pipelines with Databricks and PySpark
- Analysis and transformation with Spark SQL
- Machine learning models with Spark MLlib
- Case study: a data warehouse on Databricks and Delta Lake

Sample schedule

6 days onsite, or 12 half-day sessions online. We fit days and session lengths to your team's calendar.

Day 1	8 h	Data engineering foundations
Day 2	8 h	Data engineering foundations (continued); Apache Spark and PySpark
Day 3	8 h	Apache Spark and PySpark (continued); Databricks; The Hadoop ecosystem
Day 4	8 h	The Hadoop ecosystem (continued); Governance, visualisation and best practice
Day 5	8 h	Governance, visualisation and best practice (continued); Projects
Day 6	5 h	Projects (continued)

Assessment and certification

How learning is checked

- A short knowledge check on day one and again at the end, so you can see each person's progress
- Graded hands-on labs in every module, reviewed by the trainer
- A final exercise that applies the programme to a problem like your team's own

Certificate and exams

- Participants who complete the programme receive a Cognitive Networks AI certificate of completion.

Formats and cohort size

- Onsite: At your office or a venue you choose, in full days.
- Online, live: Instructor-led over video, usually in half-day sessions so people can keep up with work.
- Hybrid: Some days in a room together, the rest online. Useful when a team works across cities.

Cohort size: 15 to 25 people per trainer. Larger teams run as parallel cohorts on the same outline, so everyone learns the same thing.

Pricing: We quote per cohort, based on duration, format and number of cohorts. There is no per-seat price list: tell us your team and we send a proposal within one business day.

Bring this programme to your team

Request a proposal at cognitivenetworks.ai/contact?programme=data-engineering

reachus@cognitivenetworks.ai · Bangalore +91 72043 05303 · Dallas +1 (303) 827-7415