Cloud Soft Solutions

HORIZON

Multi-Cloud Data Engineering, Analytics and AI Program 2026

AWS · Azure · GCP · Python · Spark · Kafka · Airflow · dbt · Power BI · ML · GenAI · Agentic AI

■ Fresh Graduate	■ Working Professional	■ Analyst → Data Engineer	■ Developer → ML/AI Engineer

■ Python Data Stack	■■ AWS Data Cloud	■ Azure Synapse+ML	■ GCP BigQuery+Vertex	■ Data Engineering	■ Analytics + BI	■ ML + MLOps	■ GenAI + Agentic AI	■ Data Gov + Security

5500+ Placements	16 Weeks	12 Projects	500+ Hours	9 Tracks	Rs.28 LPA Highest Pkg

9 PARALLEL TRACKS · ALL FROM DAY 1 · 12 INDUSTRY REAL-TIME PROJECTS

Expected Salary: Rs.6 LPA - Rs.28 LPA | $80K - $190K (Global)

9666019191 | www.cloudsoftsol.com | Hyderabad | 2026 Batch

Cloud Soft Solutions — Hyderabad's #1 Tech Career Institute | 5500+ Placed Professionals

WHO IS THE HORIZON PROGRAM FOR?

■ Fresh Graduates	B.Tech/BCA/MCA/BSc (CS, IT, Maths, Statistics). Zero experience OK. Build 12 projects and get placed.	Structured path from Python basics to cloud data platforms · Full placement support till offer
	Software engineers, analysts, scientists, finance/operations professionals moving to data/AI/ML roles.	Fast-track for partial knowledge · Add Spark, Kafka, ML, GenAI · Target Rs.8-18 LPA salary hike
■ Analysts to Data Engineers	Excel/SQL/BI analysts wanting to move up to Data Engineering, cloud analytics, or ML Engineering.	Build pipelines with Spark+Kafka+Airflow · Migrate from SQL to cloud data warehouses + ML models
■ Developers to ML/AI	Python/Java/.NET developers wanting to specialize in ML Engineering, MLOps, GenAI, or AI architecture.	Skip basics — fast-track to MLOps+LLMOps · Add GenAI, RAG, Agentic AI to engineering skills

5500+ Successful Placements across India's top product companies, MNCs, unicorns and global remote firms. Founded by veterans from Google, Amazon and TCS.

200+ Hiring Partners — direct referrals in Hyderabad, Bangalore, Pune, Chennai and US/UK/Canada/Australia. Curriculum built from 60,000+ real JD analysis.

Flexible Batches — Weekday morning, evening and weekend options. Career team schedules all interviews. Placement support until your offer is signed.

9-TRACK PARALLEL LEARNING — ALL TRACKS FROM DAY 1

All 9 domains run simultaneously from Week 1 — mirroring how Data Engineers, ML Engineers and AI Engineers work at top companies: writing Python AND querying BigQuery AND building Spark pipelines AND training models AND deploying GenAI every single week.

TRACK	MONTH 1 Wks 1-4	MONTH 2 Wks 5-8	MONTH 3 Wks 9-12	MONTH 4 Wks 13-16
Python Data	NumPy+Pandas+SQL	Advanced Pandas+Polars	PySpark+Dask+Numba	Python AI agents+tools
AWS Data	S3+Glue+Athena Wk1	Redshift+EMR+Kinesis	SageMaker+Bedrock	Lake Formation+FinOps
Azure Data	ADF+Synapse+Blob Wk1	Databricks+Event Hubs	Azure ML+OpenAI	Fabric+Purview+Gov
GCP Data	BigQuery+GCS Wk1	Dataproc+Pub/Sub	Vertex AI+Gemini	Dataflow+Looker+Gov
Data Engineering	Airflow+dbt basics	Kafka+Spark Streaming	Delta Lake+Iceberg	Real-time+CDC+Mesh
Analytics+BI	SQL+EDA+Tableau	Power BI+Looker+Stats	A/B Testing+Forecasting	Self-serve+AI Analytics
ML+MLOps	scikit-learn+MLflow	XGBoost+PyTorch+KFP	SageMaker+Vertex MLOps	LLMOps+Model Monitor
GenAI+Agentic	Prompts+RAG basics	LangChain+LangGraph	CrewAI+Fine-tuning	Multi-Agent+MCP+VLLM
Data Gov+Sec	IAM+data masking	Data catalog+quality	GDPR+CCPA+lineage	DataMesh+Compliance

6-7 AM	DSA + SQL challenge — LeetCode (Easy to Medium) + 1 SQL query on real dataset daily
9-12 PM	PRIMARY: Data Engineering + Cloud (AWS/Azure/GCP rotating deep-dive weekly)
1-3 PM	PARALLEL: ML + Analytics + BI — all run every afternoon from Week 1
3-5 PM	Project building — real pipelines, dashboards, models, deployed to cloud
6-7 PM	GenAI + Agentic AI + Data Governance — daily 1-hour labs from Week 1
8-9 PM	GitHub commit · Kaggle notebook · Blog · LinkedIn post · Architecture doc

MONTH 1	ALL 9 TRACKS — Foundations Sprint	Python+Pandas+SQL+AWS S3+Azure ADF+BigQuery+Airflow+scikit-learn+Prompts	Weeks 1-4 | 2026

Python Data

- ■ Python OOP, comprehensions, decorators, generators
- ■ NumPy: arrays, broadcasting, vectorized operations
- ■ Pandas: DataFrames, groupby, merge, pivot, reshape
- ■ Matplotlib + Seaborn + Plotly interactive charts

AWS Data

- ■ S3: buckets, prefixes, lifecycle rules, versioning, events
- ■ AWS Glue: crawlers, catalog, ETL jobs (Python shell)
- ■ Athena: query S3 data with SQL, partitioning, formats
- ■ CloudWatch metrics + Cost Explorer + IAM for data

Azure Data

- ■ Azure Data Factory: linked services, datasets, pipelines
- ■ Azure Blob + Data Lake Gen2: directories, ACLs, SAS tokens
- ■ Azure SQL Database + Azure Cosmos DB basics
- ■ Synapse Analytics: SQL pools, Spark pools, Link

GCP Data

- ■ BigQuery: datasets, tables, partitioning, clustering
- ■ GCS: buckets, object storage, lifecycle, signed URLs
- ■ Cloud Composer (Airflow managed): first DAG on GCP
- ■ BigQuery SQL: analytic functions, nested records, UNNEST

Data Engineering

- ■ Apache Airflow: DAGs, operators, sensors, connections
- ■ dbt: models, sources, tests, documentation, macros
- ■ EDA-to-pipeline: from Jupyter exploration to Airflow DAG
- ■ Data pipeline design patterns: incremental vs full refresh

Analytics+BI

- ■ SQL advanced: window functions, CTEs, recursive queries
- ■ Tableau: connect to cloud, calculated fields, dashboards
- ■ Power BI: Power Query, DAX basics, publish to service
- ■ Statistical EDA: distributions, outliers, correlation

ML+MLOps

- ■ scikit-learn: preprocessing, pipelines, cross-validation
- ■ RandomForest + XGBoost + LightGBM for tabular data
- ■ MLflow: experiment tracking, metrics, model registry
- ■ FastAPI model serving + Pydantic data validation

GenAI+Agentic

- ■ LLM fundamentals: tokens, context window, temperature
- ■ OpenAI + Anthropic API: GPT-4o calls, structured output
- ■ Prompt engineering: zero-shot, few-shot, CoT, JSON mode
- ■ LangChain basics: chains, memory, first RAG prototype

Data Gov+Sec

- ■ Data classification: PII, sensitive, internal, public
- ■ AWS IAM + column-level security in Redshift/Athena
- ■ Data masking: pandas-based PII redaction pipelines
- ■ Great Expectations: data quality checks in Airflow DAGs

Week-by-Week Plan — Month 1

Week 1	Python Core (OOP, comprehensions, NumPy, Pandas) · AWS: S3+Glue+Athena first queries · Azure: ADF pipeline (Blob to Azure SQL) · BigQuery: first queries on public datasets · Airflow: install + first DAG · ML: scikit-learn first model (RandomForest on Titanic) · GenAI: OpenAI API + 10 prompt experiments · Analytics: Tableau + Power BI connect to cloud · Data Gov: IAM setup + Great Expectations first checkpoint · DSA: 8 Easy SQL
Week 2	Python advanced (generators, decorators) + Pandas EDA deep dive · AWS: Glue crawler + Athena partitioned table + cost estimation · Azure: Synapse SQL pool + Spark pool first notebook · BigQuery: analytic functions + nested records + public datasets EDA · dbt: first models + sources + tests on a real dataset · ML: XGBoost + MLflow tracking · GenAI: LangChain LLMChain + first RAG with PDF · Analytics: Tableau dashboard published · Data Gov: data classification pipeline + PII masking · DSA: 6 Easy SQL
Week 3	Python Polars (fast DataFrames) + Seaborn/Plotly visualizations · AWS: Glue ETL job (S3 → Parquet → Athena) · Azure: Data Factory pipeline (on-premise to cloud simulation) · BigQuery: dbt models on BigQuery + Cloud Composer DAG · Airflow: advanced operators (PythonOperator, BranchOperator) · ML: LightGBM + SHAP explainability + model comparison · GenAI: LangChain tool use + first agent with Airflow monitoring tool · Analytics: Power BI DAX measures + drill-through reports · Data Gov: Great Expectations + dbt tests in Airflow · DSA: 6 Easy+2 Medium SQL
Week 4	PROJECT SPRINT: Build Projects 1 and 2 · Full Airflow+dbt pipeline runs · Glue+Athena+BigQuery all queried · ML model in MLflow registry · GenAI chatbot integrated · Tableau+Power BI dashboards published · LeetCode: 26 Easy + 10 SQL

MONTH 1 — REAL-TIME PROJECTS

PRJ 01	Multi-Cloud Data Lakehouse Foundation (AWS+Azure+GCP)	Stack: Python, AWS S3+Glue+Athena+Lake Formation, Azure Data Lake Gen2+ADF+Synapse, BigQuery+GCS, Apache Airflow, dbt, Delta Lake
Description	Production-grade data lakehouse across 3 clouds: AWS side — S3 raw/silver/gold zones, Glue crawlers catalog all assets, Athena queries with partition projection, Lake Formation column-level security; Azure side — Data Lake Gen2 + ADF pipelines + Synapse SQL pool; GCP side — BigQuery partitioned tables + GCS + Cloud Composer. Unified dbt project with cross-cloud models. Airflow orchestrates all ETL jobs. Delta Lake format ensures ACID transactions on AWS+Azure.	
Deliverables	Delta Lake on S3+ADLS · dbt cross-cloud models · Airflow multi-cloud DAG · Lake Formation policies · Glue catalog · Athena+Synapse+BigQuery querying · Architecture diagram · Cost report	
Target Roles	**Data Engineer | Cloud Data Engineer | Analytics Engineer | Platform Data Engineer**	

PRJ 02	End-to-End Analytics Platform + BI Dashboard	Stack: Python, Pandas, SQL, BigQuery, Redshift, Synapse, Tableau, Power BI, Looker Studio, Great Expectations, dbt

Description	Complete analytics platform: Python EDA pipeline cleans and profiles retail dataset (10M+ rows). dbt models transform raw data to marts (customer_360, product_performance, revenue_forecasting). BigQuery+Redshift+Synapse serve as analytical backends. Tableau dashboard (revenue, cohort, funnel, geo maps) + Power BI report (DAX measures, Row-Level Security, publish to service) + Looker Studio public report. Great Expectations ensures data quality at every layer.
Deliverables	Python EDA pipeline · dbt models (staging, int, marts) · BigQuery+Redshift+Synapse · Tableau dashboard · Power BI report with RLS · Looker Studio · Great Expectations report · dbt test suite
Target Roles	**Analytics Engineer | BI Developer | Data Analyst | Data Analyst Lead | Business Intelligence Engineer**

MONTH 2 **ALL 9 TRACKS — Intermediate Depth** **Kafka+Spark Streaming+Kinesis · Databricks · PyTorch+NLP · LangGraph · Power BI Advanced** **Weeks 5-8 | 2026**

Python Data

- ■ PySpark: DataFrames, SQL, joins, window functions, MLlib
- ■ Dask: parallel pandas for large-scale data processing
- ■ Python Kafka client (confluent-kafka): producer+consumer
- ■ Python + Great Expectations + Pandera: data validation

AWS Data

- ■ Amazon Redshift: clusters, distribution keys, sort keys, WLM
- ■ AWS EMR: Spark clusters, step functions, auto-scaling
- ■ Amazon Kinesis: Data Streams, Firehose, Data Analytics
- ■ MSK (Managed Kafka): topics, consumer groups, monitoring

Azure Data

- ■ Azure Databricks: clusters, notebooks, Delta tables, Unity Catalog
- ■ Azure Event Hubs: namespaces, consumer groups, Kafka API
- ■ Azure Stream Analytics: SQL streaming queries, windowing
- ■ Azure Synapse Spark: notebooks, linked services, Delta Lake

GCP Data

- ■ Google Dataproc: Spark on GCP, auto-scaling clusters
- ■ Google Pub/Sub: topics, subscriptions, push+pull
- ■ Dataflow (Apache Beam): batch+streaming unified pipelines
- ■ BigQuery ML: ARIMA_PLUS, XGBoost, logistic regression in SQL

Data Engineering

- ■ Apache Kafka: topics, partitions, replication, consumer groups
- ■ Spark Structured Streaming: readStream, watermarks, stateful
- ■ Apache Flink basics: event time, exactly-once, CEP
- ■ Change Data Capture (CDC): Debezium + Kafka + Sink

Analytics+BI

- ■ Advanced Power BI: composite models, aggregations, dataflows
- ■ Statistical methods: A/B testing, confidence intervals, effect size
- ■ Time series analysis: ARIMA, Prophet, ETS, seasonal decomp
- ■ dbt metrics layer + semantic layer for self-serve analytics

ML+MLOps

- ■ PyTorch: tensors, autograd, training loop, custom datasets
- ■ HuggingFace Transformers: BERT, DistilBERT text classification
- ■ MLflow Model Registry: staging, production promotion, A/B
- ■ Kubeflow Pipelines: ML workflow on Kubernetes

GenAI+Agentic

- ■ LangGraph: stateful multi-step agent workflows with memory
- ■ Vercel AI SDK: streamText, useChat, Server-Sent Events
- ■ Pinecone + Weaviate + pgvector: production vector stores
- ■ RAGAS: automated evaluation of RAG pipelines (faithfulness)

Data Gov+Sec

- ■ Apache Atlas / DataHub: metadata management + data catalog
- ■ dbt exposures + sources: document data lineage in code
- ■ GDPR Article 17: right-to-erasure pipeline implementation
- ■ Data quality SLAs: monitoring + alerting with Great Expectations

Week-by-Week Plan — Month 2

Week 5	PRIMARY: PySpark DataFrames + Spark SQL + EMR cluster · AWS: Redshift distribution keys + WLM queuing + Kinesis Streams · Azure: Databricks Delta Live Tables + Event Hubs Kafka API · GCP: Dataproc Spark + Pub/Sub + BigQuery ML first model · Data Eng: Kafka topics+partitions+consumer groups · ML: PyTorch first neural network + HuggingFace BERT text classify · GenAI: LangGraph stateful workflow + Pinecone setup · Analytics: A/B testing with scipy+statsmodels · Data Gov: DataHub install + Glue catalog integration · DSA: 8 Easy+4 Medium SQL
Week 6	PRIMARY: Dask parallel DataFrames + Python Kafka producer+consumer · AWS: EMR Spark job on S3 + MSK managed Kafka · Azure: Azure Stream Analytics SQL streaming + Databricks Unity Catalog · GCP: Dataflow Apache Beam batch+stream + Pub/Sub streaming · Data Eng: Spark Structured Streaming (readStream+watermarks) · ML: HuggingFace BERT fine-tune sentiment classification + MLflow · GenAI: Vercel AI SDK useChat + LangGraph RAG agent · Analytics: Prophet time series forecast + Power BI composite model · Data Gov: dbt lineage docs + DataHub data catalog · DSA: 5 Easy+6 Medium SQL
Week 7	PRIMARY: PySpark MLlib + Spark ML pipeline + Pandas UDFs · AWS: Kinesis Firehose to S3+Redshift + Redshift Spectrum · Azure: Synapse Spark Delta Lake + Azure ML workspace · GCP: BigQuery ML XGBoost + Looker Studio advanced · Data Eng: CDC with Debezium + Kafka Connect + Sink connector · ML: Kubeflow Pipelines first ML workflow + MLflow Registry · GenAI: RAGAS evaluation + LangChain advanced tools · Analytics: dbt metrics layer + semantic layer · Data Gov: GDPR erasure pipeline + data quality SLAs · DSA: 4 Easy+5 Medium
Week 8	PROJECT SPRINT: Build Projects 3, 4 and 5 · Kafka streaming pipeline live · Spark jobs on EMR+Dataproc+Databricks · Redshift+BigQuery ML models deployed · LangGraph agent integrated · LeetCode: 60 Easy+22 Medium

MONTH 2 — REAL-TIME PROJECTS

PRJ 03 **Real-Time Streaming Data Platform (Kafka+Spark+Flink)** **Stack: Apache Kafka, Spark Structured Streaming, Apache Flink, Kinesis, Azure Event Hubs, Pub/Sub, Python, Debezium, Delta Lake**

Description	Production real-time streaming platform: Kafka (MSK on AWS) ingests 500K events/sec from e-commerce clickstream. Spark Structured Streaming processes events with 5-second micro-batches, computes running totals and sessionization, writes to Delta Lake on S3. Flink CEP detects fraud patterns (velocity checks, geo-anomalies) in real-time. Debezium CDC captures database changes from PostgreSQL → Kafka → downstream consumers. Azure Event Hubs mirrors for Azure-side analytics. Pub/Sub feeds BigQuery Streaming Insert.
Deliverables	Kafka cluster + 5 topics · Spark Streaming + Delta Lake · Flink CEP fraud rules · Debezium CDC pipeline · Event Hubs + Stream Analytics · Pub/Sub + BigQuery Streaming · Grafana metrics dashboard · Architecture diagram
Target Roles	**Data Engineer | Streaming Engineer | Real-time Analytics Engineer | Platform Engineer**

PRJ 04 — Cloud Data Warehouse + Lakehouse Analytics (Redshift+Synapse+BigQuery)

Stack: Redshift, Azure Synapse, BigQuery, dbt, Airflow, Apache Iceberg, Delta Lake, Tableau, Power BI, Looker

Description	Enterprise multi-cloud data warehouse: Redshift (AWS) — fact/dim star schema with distribution keys, Spectrum for S3 external tables. Azure Synapse — dedicated SQL pool with columnstore indexes, Synapse Pipelines. BigQuery — partitioned+clustered tables, authorized views for RLS. dbt project with 50+ models across all 3 warehouses (staging, intermediate, marts). Airflow orchestrates daily loads. Apache Iceberg handles time-travel and schema evolution. Unified semantic layer via dbt metrics.
Deliverables	Redshift + Synapse + BigQuery models · dbt 50+ models · Apache Iceberg time-travel · Airflow orchestration · Tableau + Power BI + Looker connected · Data quality test suite · dbt lineage documentation · Cost comparison report
Target Roles	**Data Warehouse Engineer | Analytics Engineer | dbt Developer | BI Developer**

PRJ 05 — Machine Learning Pipeline + MLOps on 3 Clouds

Stack: PyTorch, HuggingFace, scikit-learn, XGBoost, MLflow, Kubeflow Pipelines, SageMaker Pipelines, Vertex AI Pipelines, Seldon Core, Evidently

Description	End-to-end MLOps: Kubeflow Pipeline on EKS — data ingestion → feature engineering → XGBoost + PyTorch model training → MLflow experiment comparison → model registry → Seldon Core canary deployment (10% traffic). SageMaker Pipeline for AWS-managed MLOps. Vertex AI Pipeline for GCP. HuggingFace BERT fine-tuned for sentiment. Evidently monitors data drift weekly and triggers retraining. Cross-cloud model registry in MLflow.
Deliverables	Kubeflow + SageMaker + Vertex AI pipelines · MLflow cross-cloud registry · Seldon Core canary · Evidently drift reports · HuggingFace BERT fine-tuned · PyTorch + XGBoost comparison · 3-cloud MLOps architecture · Documentation
Target Roles	**MLOps Engineer | ML Engineer | ML Platform Engineer | AI Infrastructure Engineer**

MONTH 3 — ALL 9 TRACKS — Advanced Engineering

SageMaker+Vertex AI · Azure ML+OpenAI · Data Mesh · LLMOps · Feature Store · Forecast+Optimization

Weeks 9-12 | 2026

Python Data

- ■ Python ONNX Runtime: model inference without heavy ML libraries
- ■ Polars + DuckDB: blazing-fast in-process analytics in Python
- ■ Python Ray: distributed computing for large-scale ML training
- ■ Python asyncio+aiohttp: async data ingestion pipelines

AWS Data

- ■ SageMaker: Feature Store, Model Monitor, Clarify (bias+explain)
- ■ AWS Bedrock: Claude, Llama 3, Titan via Python boto3 SDK
- ■ Bedrock Knowledge Bases: managed RAG + OpenSearch Serverless
- ■ Lake Formation + Macie + AWS Glue DataBrew: data governance

Azure Data

- ■ Azure ML: AutoML, Responsible AI, model explanations dashboard
- ■ Azure OpenAI Service: GPT-4o, embeddings, DALL-E via Python
- ■ Microsoft Fabric: OneLake, Lakehouse, Notebooks, Data Factory
- ■ Microsoft Purview: data catalog, lineage, classification policies

GCP Data

- ■ Vertex AI: custom training, Feature Store, Pipelines, Matching Engine
- ■ Gemini API: multimodal (text+image+video) via Python SDK
- ■ Looker: LookML models, explores, dashboards, Looker API
- ■ BigQuery Analytics Hub + Dataplex governance platform

Data Engineering

- ■ Delta Lake advanced: OPTIMIZE, VACUUM, Z-ORDER, CDF
- ■ Apache Iceberg: hidden partitioning, branching, compaction
- ■ Data Mesh: domain ownership, data products, federated governance
- ■ dbt advanced: macros, packages, snapshots, incremental models

Analytics+BI

- ■ Advanced forecasting: Prophet + LSTM + Transformer time series
- ■ Causal inference: A/B testing, diff-in-diff, propensity scoring
- ■ Optimization: linear programming, simulation with Python scipy
- ■ Embedded analytics: Tableau Embedded + Superset + Metabase

ML+MLOps

- ■ Feast: online + offline feature store serving and management
- ■ Model explainability: SHAP TreeExplainer, LIME, integrated gradients
- ■ LLMOps: prompt versioning, evaluation, cost tracking (LangFuse)
- ■ Ray Serve: scalable model serving with batching on Kubernetes

GenAI+Agentic

- ■ CrewAI: role-based multi-agent for data analysis automation
- ■ LangChain advanced: multi-vector retrieval, reranking (Cohere)
- ■ Fine-tuning with QLoRA + Unsloth on Colab A100
- ■ Multimodal RAG: text + image + tabular data in one pipeline

Data Gov+Sec

- ■ Data lineage: OpenLineage + Marquez across Airflow+Spark+dbt
- ■ CCPA/GDPR compliance: consent management + data inventory
- ■ Row-Level Security: BigQuery, Redshift, Power BI, Tableau
- ■ Data contracts: defining and enforcing schema agreements

Week-by-Week Plan — Month 3

Week 9	PRIMARY: SageMaker Feature Store + Model Monitor + Clarify bias detection · AWS: Bedrock Knowledge Bases + Bedrock Claude python SDK · Azure: Azure ML AutoML + Responsible AI dashboard · GCP: Vertex AI custom training + Feature Store + Gemini API · Data Eng: Delta Lake advanced (OPTIMIZE+Z-ORDER+CDF) + Iceberg hidden partitioning · ML: Feast feature store + SHAP explainability · GenAI: CrewAI data analysis agents · Analytics: Prophet LSTM hybrid forecasting · Data Gov: OpenLineage+Marquez lineage tracking · DSA: 7 Medium+1 Hard
Week 10	PRIMARY: Python DuckDB+Polars in-process analytics + Ray distributed training · AWS: Lake Formation row/column-level policies + Glue DataBrew · Azure: Microsoft Fabric OneLake + Microsoft Purview catalog · GCP: Looker LookML + BigQuery Analytics Hub + Dataplex · Data Eng: Data Mesh patterns + dbt advanced macros+snapshots · ML: LLMOps with LangFuse + Ray Serve model serving · GenAI: LangChain multi-vector retrieval + Cohere reranking · Analytics: Causal inference A/B testing + propensity scoring · Data Gov: GDPR/CCPA compliance pipeline + data contracts · DSA: 5 Medium+2 Hard
Week 11	PRIMARY: ONNX Runtime + asyncio async ingestion pipelines · AWS: SageMaker Ground Truth labeling + Batch Transform · Azure: Azure OpenAI embeddings + Azure ML Pipelines advanced · GCP: Vertex AI Matching Engine (vector similarity at scale) · Data Eng: Advanced Iceberg + multi-cloud Delta sharing · ML: Multimodal models (text+image) + model registry governance · GenAI: Multimodal RAG + QLoRA fine-tuning on Colab · Analytics: Embedded analytics + Superset+Metabase setup · Data Gov: Row-Level Security all platforms + data catalog finalization · DSA: System design (data platform, ML system, streaming)
Week 12	PROJECT SPRINT: Build Projects 6 and 7 · SageMaker+Vertex AI Feature Stores live · Purview+Dataplex governance · LangFuse LLMOps tracking · CrewAI data agents deployed · LeetCode: 80 Easy+36 Medium+3 Hard

MONTH 3 — REAL-TIME PROJECTS

PRJ 06	GenAI Data Assistant — Enterprise Knowledge Query Platform	Stack: LangChain, CrewAI, GPT-4o, Pinecone, Weaviate, BigQuery, Redshift, Azure OpenAI, RAGAS, LangFuse, Python, NestJS
Description	Enterprise GenAI analytics assistant: users query data in natural language. LangChain SQL agent generates and executes SQL against BigQuery/Redshift/Synapse. LangChain RAG pipeline answers questions from documents (PDF annual reports, Confluence wikis, Slack threads) via Pinecone+Weaviate hybrid search. CrewAI Data Analyst agent interprets results and generates chart code. Azure OpenAI fallback. LangFuse tracks all LLM calls (cost+quality). RAGAS automated quality eval.	
Deliverables	LangChain SQL agent · RAG document pipeline · CrewAI analyst agent · Pinecone+Weaviate hybrid search · BigQuery+Redshift+Synapse connectors · LangFuse LLMOps · RAGAS evaluation · Azure OpenAI fallback · Usage analytics dashboard	
Target Roles	**AI/ML Engineer | GenAI Engineer | Data & AI Engineer | LLM Engineer | Data Product Engineer**	

PRJ 07	Advanced ML Platform + Feature Store + Model Monitoring	Stack: Feast, SageMaker Feature Store, Vertex AI Feature Store, MLflow, Ray Serve, Evidently, Seldon, SHAP, LangFuse, Prometheus
Description	Production ML platform: Feast feature store serves 200+ features online (Redis) and offline (S3/BigQuery). SageMaker and Vertex AI Feature Stores for cloud-managed serving. Ray Serve handles high-throughput model serving with dynamic batching. MLflow Model Registry tracks 15+ model versions across 3 experiments. Evidently monitors 5 data drift metrics weekly. SHAP explains every prediction. Seldon Core canary deployment. Prometheus+Grafana MLOps dashboard shows latency, throughput, drift score, prediction distribution.	
Deliverables	Feast feature store (200+ features) · SageMaker+Vertex Feature Stores · Ray Serve serving · MLflow Registry · Evidently monitoring · SHAP explanations · Seldon canary · Prometheus+Grafana ML dashboard · Feature documentation	
Target Roles	**ML Platform Engineer | MLOps Engineer | Feature Engineer | AI Infrastructure Engineer**	

MONTH 4 — **ALL 9 TRACKS — Production + Capstone** — Agentic Pipelines · Multi-Cloud Governance · LLMOps · DataMesh · Real-time AI + Capstone — Weeks 13-16 | 2026

Python Data

- ■ Python operator SDK for Airflow: custom operators+sensors
- ■ Python LangGraph tools: data pipeline management agents
- ■ Python streaming inference: ONNX+TorchServe+Triton
- ■ Python data apps: Streamlit+Gradio+Panel production apps

AWS Data

- ■ SageMaker JumpStart: foundation models + fine-tuning
- ■ AWS Bedrock Agents: build agents with knowledge bases
- ■ AWS Glue 4.0: streaming ETL, Ray, interactive sessions
- ■ FinOps for data: S3 Intelligent-Tiering, Redshift Reserved

Azure Data

- ■ Azure AI Search: vector search, semantic ranking, hybrid
- ■ Microsoft Fabric: Direct Lake mode, shortcuts, mirroring
- ■ Azure Databricks + Unity Catalog: fine-grained governance
- ■ Azure AI Studio: prompt flow, model catalog, fine-tuning

GCP Data

- ■ BigQuery Omni: query AWS+Azure data from BigQuery
- ■ Vertex AI Agent Builder: RAG apps with grounding
- ■ Gemini for BigQuery: NL2SQL + data insights + explanation
- ■ Looker Embedded: white-label analytics in your SaaS product

Data Engineering

- ■ Data Mesh implementation: domain teams + data products
- ■ Apache Hudi: upserts, incremental queries, multi-table transactions
- ■ dbt Core + dbt Cloud: CI/CD for analytics code
- ■ DataOps: version control, testing, CI/CD for data pipelines

Analytics+BI

- ■ ML-augmented analytics: anomaly detection in dashboards
- ■ NL2SQL: natural language to SQL with LLM integration
- ■ Self-serve analytics: Superset, Metabase, Evidence.dev
- ■ Product analytics: Mixpanel, Amplitude patterns in BigQuery

ML+MLOps	GenAI+Agentic	Data Gov+Sec
■ LLM fine-tuning: SFT + DPO + RLHF on domain data	■ Agentic data pipelines: LangGraph orchestrates ETL+ML tasks	■ Data product contracts: schema versioning + SLA enforcement
■ Model governance: access control, audit logs, responsible AI	■ MCP servers: expose BigQuery+S3+pipelines to AI agents	■ Privacy engineering: differential privacy + k-anonymity
■ Batch+online inference: Spark ML + real-time REST endpoint	■ Multimodal GenAI: GPT-4o Vision analyzing charts and images	■ Cloud compliance: SOC2 for data platforms + ISO27001
■ Multi-cloud model serving: SageMaker+Vertex AI+Azure ML	■ GenAI for data quality: LLM-based anomaly explanation	■ DataHub + Purview + Dataplex: unified governance at scale

Week-by-Week Plan — Month 4

Week 13	PRIMARY: LangGraph data pipeline agent + MCP server for data tools · AWS: SageMaker JumpStart fine-tuning + Bedrock Agents + FinOps · Azure: Azure AI Studio prompt flow + Fabric Direct Lake + Unity Catalog · GCP: BigQuery Omni + Vertex AI Agent Builder + Gemini NL2SQL · Data Eng: Data Mesh domain teams + Apache Hudi upserts + DataOps CI/CD · ML: LLM fine-tuning (SFT+DPO) + multi-cloud serving · GenAI: Agentic data pipeline + MCP BigQuery tool · Analytics: NL2SQL implementation + anomaly detection in dashboards · Data Gov: Data product contracts + privacy engineering · DSA: Mock OA x2
Week 14	PRIMARY: Python streaming inference (Triton+ONNX) + Streamlit production app · AWS: Glue 4.0 streaming ETL + Redshift FinOps optimization · Azure: Azure AI Search hybrid + Fabric mirroring + Databricks Unity · GCP: Looker Embedded + BigQuery ML advanced + Gemini for BigQuery · Data Eng: dbt Cloud CI/CD + Apache Hudi advanced + DataOps · ML: Model governance audit + responsible AI report + batch inference · GenAI: Multimodal RAG (text+image+chart) + MCP multi-tool server · Analytics: Superset+Metabase self-serve + product analytics patterns · Data Gov: SOC2 evidence + DataHub+Purview+Dataplex unified governance · DSA: Mock OA x2
Week 15	CAPSTONE SPRINT Week 1: Architect + build Project 12 · Multi-cloud data platform provisioned · Streaming + batch pipelines running · ML models on 3 clouds · GenAI analytics assistant live · Governance framework applied · FinOps optimization implemented
Week 16	CAPSTONE SPRINT Week 2+PORTFOLIO: Polish + deploy all 12 projects · Architecture diagrams + data flow diagrams · Kaggle notebook · Blog posts · Update GitHub+LinkedIn+Resume · 5 mock technical+3 system design+2 HR behavioral interviews · LeetCode: 90 Easy+45 Medium+5 Hard COMPLETE

MONTH 4 — REAL-TIME PROJECTS

PRJ 08	Agentic Data Pipeline Orchestrator	Stack: LangGraph, CrewAI, Python, MCP, BigQuery, Redshift, Airflow, AWS Bedrock, GPT-4o, LangFuse, Great Expectations
Description	AI agent system that autonomously monitors, diagnoses and repairs data pipelines. LangGraph workflow: Pipeline Monitor Agent (polls Airflow API + Prometheus) detects failed DAGs, Data Quality Agent (runs Great Expectations checks on new partitions) finds anomalies, Root Cause Agent (queries logs+metrics+dbt docs via LLM) diagnoses issues, Repair Agent (restarts tasks, applies dbt fixes, creates GitHub issue). Custom MCP server exposes BigQuery+Redshift+Airflow APIs to agents. LangFuse traces all agent decisions. Evaluated on 30 real pipeline failure scenarios.	
Deliverables	LangGraph pipeline monitor workflow · Great Expectations integration · MCP server (BigQuery+Airflow tools) · AWS Bedrock reasoning · LangFuse observability · 30-scenario evaluation · Auto-repair demo video · Pipeline health dashboard	
Target Roles	**Agentic AI Engineer | Data Platform Engineer | DataOps Engineer | AI Infrastructure Engineer**	

PRJ 09	Real-Time Fraud Detection + ML Scoring Engine	Stack: Kafka, Spark Streaming, Flink CEP, XGBoost, Feast, MLflow, Ray Serve, Kinesis, Event Hubs, BigQuery, Grafana
Description	Production fraud detection: Kafka ingests 200K transactions/second. Flink CEP detects 15+ fraud patterns (velocity, geo-anomaly, device-fingerprint) in real-time (<100ms). XGBoost model (AUC 0.99+) served via Ray Serve with <30ms P99 latency. Feast serves 50+ pre-computed features from Redis online store. Batch Spark job retrains model daily on EMR, registers in MLflow, promotes via Seldon canary. Evidently monitors feature drift. Grafana dashboard shows fraud rate, model score distribution, feature importance.	
Deliverables	Kafka+Flink CEP streaming · XGBoost AUC 0.99+ · Ray Serve <30ms latency · Feast online store · MLflow model registry · Seldon canary deployment · Evidently drift monitoring · Grafana dashboard · Business impact report	
Target Roles	**ML Engineer | Streaming Engineer | Data Engineer | FinTech ML Engineer | Real-time AI Engineer**	

PRJ 10	NLP Analytics Platform + Text Intelligence Suite	Stack: HuggingFace, BERT, PyTorch, SpaCy, LangChain, BigQuery, Redshift, Airflow, Streamlit, MLflow, SageMaker
Description	Production NLP platform on e-commerce reviews (5M+ reviews). Pipeline: Airflow ingests reviews daily → SpaCy NER extracts entities (product, brand, issue) → BERT fine-tuned (PyTorch) for sentiment+category classification (92% accuracy) → topic modeling (BERTopic) discovers themes → Aspect-Based Sentiment (ABSA) scores product features. All stored in BigQuery. Streamlit dashboard shows real-time sentiment trends, brand health, competitive intelligence. SageMaker Batch Transform for large-scale inference.	
Deliverables	BERT fine-tuned classifier (92% acc) · SpaCy NER pipeline · BERTopic topic modeling · ABSA feature scoring · Airflow daily pipeline · BigQuery analytics store · Streamlit live dashboard · MLflow registry · SageMaker Batch Transform	
Target Roles	**NLP Engineer | ML Engineer | Data Scientist | Analytics Engineer | AI Product Engineer**	

PRJ 11	IoT Data Platform + Predictive Maintenance AI	Stack: Python, Kafka, Spark Streaming, InfluxDB, TimescaleDB, AWS IoT Core, Azure IoT Hub, GCP IoT, Flink, Prophet, LSTM, Grafana

Description	End-to-end IoT data platform: 10,000 simulated sensors publish telemetry via AWS IoT Core + Azure IoT Hub + GCP IoT. Kafka aggregates all streams. Spark Streaming computes sliding window anomalies and writes to InfluxDB (real-time) + TimescaleDB (historical). Flink CEP detects equipment failure precursors. Prophet+LSTM ensemble predicts maintenance windows 24h ahead (accuracy 88%+). Grafana shows live sensor metrics + predictive maintenance calendar. Auto-alerts via PagerDuty when failure probability >85%.
Deliverables	10K sensor simulation · AWS IoT+Azure IoT+GCP IoT · Kafka ingestion · Spark Streaming anomalies · InfluxDB+TimescaleDB · Flink CEP fault detection · Prophet+LSTM predictive model (88% acc) · Grafana live dashboard · PagerDuty alerts
Target Roles	**IoT Data Engineer | Streaming Engineer | ML Engineer | Data Platform Engineer**

CAPSTONE — PROJECT 12: ALL 9 TRACKS INTEGRATED

PROJECT 12	Unified Multi-Cloud AI Data Platform — End-to-End Production System		CAPSTONE · 9 Tracks · 4 Months · Production
Python+Data Eng	PySpark+Kafka+Airflow+dbt+Delta Lake+Iceberg+Feast+Ray	AWS+Azure+GCP	S3+Redshift+SageMaker+ADLS+Synapse+Azure ML+BigQuery+Vertex AI
Analytics+BI	dbt+BigQuery ML+Tableau+Power BI+Looker+Superset+NL2SQL	ML+MLOps	XGBoost+PyTorch+HuggingFace+MLflow+Kubeflow +Seldon+Evidently
GenAI+Agentic	LangChain+LangGraph+CrewAI+MCP+VLLM+LangFuse+Pinecone+RAGAS	Gov+Security	DataHub+Purview+Dataplex+OpenLineage+Great Expectations+SOC2

What it is	A production e-commerce intelligence platform: real-time streaming, multi-cloud data lakehouse, ML-powered personalization, GenAI analytics assistant, agentic pipeline orchestrator, and enterprise governance — all integrated across AWS+Azure+GCP.
Data Platform	Kafka ingests clickstream + transactions + IoT sensors → Spark Streaming (real-time) + Airflow+dbt batch (T+1) → Delta Lake on S3/ADLS/GCS (lakehouse) → Redshift+Synapse+BigQuery (warehouses) → dbt 80+ models (staging+intermediate+marts) → Feast feature store (300+ features) → ML models served via Ray Serve+Seldon
AI+Analytics	ML: XGBoost recommendation engine + BERT sentiment + LSTM churn prediction + PyTorch demand forecast. LangGraph agentic pipeline orchestrator monitors and auto-repairs. GenAI SQL assistant (NL2SQL on BigQuery+Redshift). Tableau+Power BI+Looker dashboards. Superset self-serve analytics. GPT-4o Vision analyzes uploaded charts and images for business users.
GenAI Layer	LangChain RAG on 50K product+support documents (Pinecone hybrid search+Cohere Rerank). CrewAI analyst agents generate weekly business intelligence reports. VLLM serves fine-tuned Llama 3.2 for product descriptions. Multimodal RAG (text+image). MCP server exposes BigQuery+Redshift+Airflow APIs to AI agents. LangFuse tracks all LLM calls. RAGAS evaluates RAG quality.
Governance	DataHub+Microsoft Purview+Dataplex unified catalog. OpenLineage tracks data lineage across Airflow+Spark+dbt. Great Expectations 500+ data quality checks. Row-Level Security in BigQuery+Redshift+Power BI+Tableau. GDPR erasure pipeline. Data contracts (Soda.io) between domains. SOC2 compliance evidence collected for all data access.
Deliverables	**Live multi-cloud platform · GitHub repo+CI/CD for all pipelines · dbt 80+ models with docs · Architecture diagrams · Jupyter notebooks · Tableau+Power BI dashboards · RAGAS quality report · Data catalog · LangFuse dashboard · SOC2 compliance evidence · Loom demo walkthrough**

PLACEMENT SUPPORT + CAREER OUTCOMES

5500+ Placements	200+ Hiring Partners	98% Placement Rate	Rs.28 LPA Highest Package	60 Days Avg to Offer
Students Placed	Active Partners	Of Completers	2025 Record	First Offer

PHASE 1 Wk 13-14	Portfolio and Brand	ATS-optimized resume · GitHub profile review (repos, notebooks, READMEs, CI badges) · LinkedIn All-Star optimization · Loom demos for 12 projects · Tailored resume per role (Data Eng, ML Eng, Analytics, AI Eng, Data Scientist)
PHASE 2 Wk 15-16	Interview Preparation	10 mock interviews: 5 technical (SQL+PySpark+Python+ML) + 3 system design (data platform, ML system, real-time streaming) + 2 behavioral HR · System design: multi-cloud data platform, streaming pipeline, ML feature store · Q&A; banks: Spark, Kafka, Airflow, dbt, BigQuery, ML, GenAI
PHASE 3 Post Wk16	Job Applications	Direct referrals to 200+ hiring partners · Weekly alerts for Data/ML/AI roles · Naukri, LinkedIn, AngelList profiles optimized · Cold outreach templates for Data Engineering managers and ML leads · 90-day application tracking
PHASE 4 Ongoing	Interview Scheduling	Placement team schedules all interviews · Company briefings (tech stack, interview rounds, culture) · Post-interview debrief and targeted re-preparation · Data + AI War Room Slack channel — alumni referral pipeline
PHASE 5 Offer Stage	Offer Negotiation	Market salary benchmarking: Rs.8-28 LPA (India), $80K-190K (Global) · Counter-offer strategy without losing the offer · Offer letter review: ESOPs, joining bonus, remote allowance · Joining date negotiation and background verification support

PRODUCT GIANTS	Google, Microsoft, Amazon, Meta, Adobe, Salesforce, Uber, Stripe, Airbnb, Netflix, LinkedIn
DATA PLATFORMS	Databricks, Snowflake, dbt Labs, Fivetran, Airbyte, Monte Carlo, Atlan, Select Star
INDIAN UNICORNS	Flipkart, Razorpay, Swiggy, Zomato, CRED, PhonePe, Meesho, Zepto, Groww, Paytm
GLOBAL MNCs	McKinsey QuantumBlack, Accenture AI, Deloitte Analytics, IBM Data, TCS Digitate, Infosys Cobalt
AI/ML FIRMS	DataRobot, H2O.ai, C3.ai, Scale AI, Weights and Biases, LangChain Inc, Cohere, Hugging Face
GLOBAL REMOTE	Toptal Data track, Turing, Arc.dev Data track + 60+ US/UK/Canada/Australia remote-first data startups

JOB ROLE	KEY SKILLS	INDIA 2026	GLOBAL
Data Engineer	PySpark, Kafka, Airflow, dbt, S3, Redshift, BigQuery	Rs.6-14 LPA	$80-120K
Senior Data Engineer	Multi-cloud data platforms, Data Mesh, Streaming, IaC	Rs.12-22 LPA	$110-155K
Analytics Engineer	dbt, SQL, BigQuery/Redshift, Tableau/Power BI, Python	Rs.6-14 LPA	$80-120K
Data Analyst	SQL, Python, Tableau, Power BI, A/B testing, Statistics	Rs.4-10 LPA	$60-95K
ML Engineer	PyTorch, XGBoost, MLflow, SageMaker, Vertex AI, Ray Serve	Rs.10-22 LPA	$105-160K
MLOps Engineer	Kubeflow, MLflow, Feast, Seldon, Evidently, Prometheus	Rs.10-24 LPA	$110-165K
Data Scientist	Python, ML, Statistics, Causal Inference, A/B Testing, NLP	Rs.8-18 LPA	$100-150K
AI/ML Architect	Multi-cloud ML platforms, LLMOps, Feature Platforms	Rs.18-36 LPA	$140-190K
LLM/GenAI Engineer	LangChain, RAG, Fine-tuning, Agents, Vector DBs	Rs.12-28 LPA	$120-180K
Agentic AI Data Engineer	LangGraph, MCP, CrewAI, pipeline agents, VLLM	Rs.15-32 LPA	$130-200K
Data Platform Engineer	Multi-cloud IaC, Data Mesh, DataHub, Governance	Rs.12-24 LPA	$110-160K

Cloud Data Architect	AWS+Azure+GCP data architecture, Well-Architected	Rs.20-40 LPA	$150-200K

CERTIFICATION	PROVIDER	PRIORITY
AWS Certified Data Engineer – Associate (DEA-C01)	AWS	Very High — Data roles
AWS Certified ML Engineer – Associate (MLA-C01)	AWS	Very High — ML roles
AWS Certified Solutions Architect – Associate (SAA-C03)	AWS	Very High — Cloud
Google Professional Data Engineer	GCP	Very High — GCP Data roles
Google Professional ML Engineer	GCP	Very High — ML+Vertex AI
Microsoft DP-203: Data Engineering on Azure	Microsoft	Very High — Azure Data
Microsoft AI-900: Azure AI Fundamentals	Microsoft	High — Azure AI entry
Databricks Certified Associate Developer for Apache Spark	Databricks	Very High — Spark
Databricks Certified ML Associate	Databricks	High — Databricks ML
dbt Analytics Engineering Certification	dbt Labs	High — Analytics Engineer
Snowflake SnowPro Core Certification	Snowflake	High — Data Warehouse
DeepLearning.AI MLOps Specialization	Coursera	High — MLOps roles

WEEK 16 SUCCESS CHECKLIST + 8 RULES TO GET HIRED

OK	12 Projects Live	Deployed with live endpoints, GitHub repos, architecture diagrams, README walkthroughs, Loom demos
OK	Kaggle Profile	3+ notebooks published with 100+ votes · public datasets · competition participation
OK	GitHub Profile	All 12 projects pinned, dbt docs deployed, Airflow DAGs documented, CI/CD badges
OK	LinkedIn All-Star	All 9 tracks listed, project posts with demo videos, 300+ connections, OpenToWork
OK	Architecture Diagrams	Data flow diagrams for all major projects — essential for Data Engineering interviews
OK	ATS Resume	1-page PDF, quantified achievements (data volumes, latency, accuracy), tailored per role
OK	3 Certifications	AWS Data Engineer + Google Data Engineer + Databricks Spark minimum by Week 16
OK	dbt Documentation	Published dbt docs site with lineage graphs — unique differentiator in analytics interviews
OK	ML Report	Model comparison report (accuracy, AUC, latency) + SHAP explanation charts
OK	GenAI Demo	Working RAG chatbot + Agentic pipeline demo — most sought-after skill in 2026
OK	Mock Interviews	10 done: 5 SQL+PySpark+Python technical + 3 system design + 2 behavioral HR
OK	Tech Blog	5+ articles: PySpark optimization, dbt patterns, Kafka setup, LangChain RAG, MLOps

1	Data at Scale	Think in billions of rows, not thousands. Always design for 100x data volume from Day 1.
2	SQL is Non-Negotiable	Every data/ML/AI role requires expert SQL. 5 LeetCode SQL problems minimum per week.
3	Build in Public	Publish Kaggle notebooks, GitHub repos, dbt docs. Recruiters find you. You do not find them.
4	Cloud is the Job	Every data role is a cloud data role in 2026. AWS Data + Azure Data + BigQuery — master all 3.
5	AI Augments Data Work	LangGraph data agents, NL2SQL, GenAI for data quality — add AI to every pipeline you build.
6	Specialize by Month 4	Data Engineer OR ML Engineer OR Analytics Engineer OR GenAI Engineer. Pick one primary role.
7	Get Certified Early	AWS Data Engineer + Google Data Engineer by Month 3. Certs unlock ATS filters instantly.
8	Measure Everything	Quantify every project: processed 500GB daily, reduced query cost 60%, improved AUC from 0.82 to 0.97.

HORIZON 2026

Python · AWS · Azure · GCP · Spark · Kafka · Airflow · dbt · ML · GenAI · Agentic AI

9666019191	www.cloudsoftsol.com	Hyderabad	2026 Batch — Enroll Now

Cloud Soft Solutions — Hyderabad's Number 1 Tech Career Transformation Institute
HORIZON 2026 | 5500+ Placements | AWS · Azure · GCP · Data Engineering · Analytics · ML · GenAI

Data is the new oil. But only engineers who can refine it and make it intelligent will own the future.