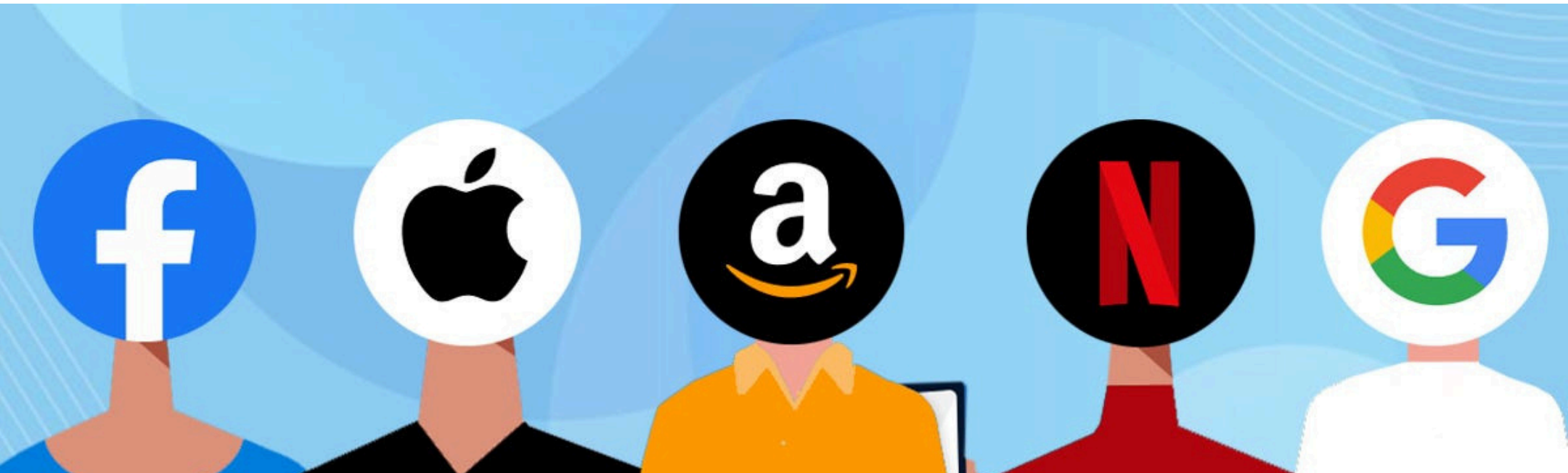
Meritshot
EDUCATION

Data Engineering Roadmap for

Working Professionals

Who Dream To Switch



Meritshot
EDUCATION

Roadmap to Land Data Engineering Roles (for Working Professionals)

How to Use This Plan

- Pick a cloud: I'll use AWS in examples (S3, EMR/Glue, MSK, Lambda, ECS/EKS).
  - Azure: ADLS, Databricks/Synapse, Event Hubs, Functions, AKS
  - GCP: GCS, Dataproc/Dataflow, Pub/Sub, Cloud Functions, GKE
- Pick a compute: PySpark for scale; SQL everywhere; dbt for analytics engineering flavor.
- Timebox: Weekdays = learning (1–1.5h/day), Weekends = project builds (3–5h).
- Deliverables over notes: Every week ends with a tangible artifact (repo, DAG, dashboard, runbook).

Meritshot
EDUCATION

Skills You Must Show as a working Professional

1. SQL mastery (window functions, joins, optimization, explain plans)
2. PySpark production patterns (partitioning, bucketing, skew handling, broadcast joins, UDF vs built-ins)
3. Batch pipelines (Airflow/Managed orchestrator + incremental loads + DQ tests)
4. Streaming (Kafka → Spark/Flink/Structured Streaming; exactly-once, watermarking, late data)
5. Lakehouse (Delta/Iceberg/Hudi; schema evolution; OPTIMIZE/Z-ORDER)
6. Cloud infra (storage, compute, IAM, costs, networking basics; containers)
7. Testing & observability (unit + data tests, lineage, metrics, logs, alerts)
8. Data modeling (3NF, Kimball, SCD, CDC; serving patterns)
9. Security (PII minimization, encryption, secrets, RBAC)
10. System design for data (batch vs streaming, SLAs/SLOs, backpressure, idempotency)

Meritshot
EDUCATION

Phase 1 — Foundations Done Right (Weeks 1–6)

Week 1 — SQL I (Joins & Filtering)

- Topics: INNER/LEFT/RIGHT/FULL, semi/anti joins (EXISTS/NOT EXISTS), NULL semantics.
- Deliverable: 20 query katas on a real dataset (e.g., NYC taxi). Add explain plan screenshots + notes.

Week 2 — SQL II (Windows & Aggregations)

- Topics: ROW_NUMBER, RANK, DENSE_RANK, LAG/LEAD, moving average, cumulative sums, gaps & islands.
- Deliverable: A “Top-K per group” report and a sessionization query with tests.

Week 3 — Python for DE

- Topics: packaging, virtualenv/poetry, type hints, logging, retries, decorators, data classes.
- Deliverable: Utility library (/lib/) with S3 helpers, logging config, and a CLI scaffold.

Meritshot
EDUCATION

Week 4 — PySpark I (Dataframes & Performance)

- Topics: transformations/actions, wide vs narrow, partitioning, bucketing, file sizing, Catalyst basics.
- Deliverable: Notebook + script that converts CSV → Parquet with partition strategy, proves shuffle reduction.

Week 5 — PySpark II (Joins & Skew)

- Topics: broadcast/hash/sort-merge joins, salting, repartitionByRange, UDF vs built-ins.
- Deliverable: Benchmark report: naive join vs broadcast vs salted; include shuffle read/write metrics.

Week 6 — Data Modeling

- Topics: OLTP vs OLAP, 3NF, Kimball star/snowflake, SCD Types 1/2, CDC patterns.
- Deliverable: ERD + star schema for a retail domain; SQL scripts for SCD2 upserts.

Milestone A (end of Week 6)

- Capability badge: Can write/read any medium-complex SQL; design a basic star schema; write a small PySpark job with optimal join & partitions.

Meritshot
EDUCATION

Phase 2 — Batch Pipelines & Lakehouse (Weeks 7–14)

Week 7 — Orchestration (Airflow)

- Topics: DAGs, Operators, Sensors, XComs, retries, SLAs; @task API.
- Deliverable: Airflow docker-compose + daily DAG that ingests raw → bronze (Parquet).

Week 8 — Lakehouse (Delta/Iceberg/Hudi)

- Topics: ACID on data lakes, OPTIMIZE, Z-ORDER, VACUUM, schema evolution.
- Deliverable: Convert bronze → silver (cleaned), use Delta with partitioning + Z-ORDER; measure query latencies.

Week 9 — Incremental Loads & CDC

- Topics: watermarking for batch, merge/upsert, idempotency keys.
- Deliverable: Incremental DAG: loads only new data, MERGE INTO Delta, with failure-safe retries.

Meritshot
EDUCATION

Week 10 — Data Quality & Lineage

- Topics: Great Expectations/dbt tests, null/duplicate checks, referential integrity, schema drift alerts.
- Deliverable: DQ suite + data docs published; failing test breaks the DAG.

Week 11 — Serving Layer Options

- Topics: Snowflake/BigQuery/Athena/Databricks SQL; materialized views; warehouse cost controls.
- Deliverable: BI-ready gold model + few performance benchmarks (cold vs warm).

Week 12 — Cost Awareness

- Topics: file size tuning, compaction, caching, lifecycle policies, spot instances.
- Deliverable: Cost experiment log: show 30–60% cost improvement with compaction + pruning.

Meritshot
EDUCATION

Week 13–14 — Batch Project #1 (End-to-End)

- Build a Retail Analytics pipeline: raw → bronze/silver/gold, SCD2 for products, DQ tests, Airflow DAGs, docs, and a simple dashboard (QuickSight/Power BI/Looker Studio).

- **KPIs:**
  - P95 job time < 15 min for sample dataset
  - DQ pass rate ≥ 99%
  - Cost/day < $1–2 (lab scale)

Milestone B

- A production-like batch project in your repo with README, runbook, and cost notes.

Meritshot
EDUCATION

Phase 3 — Streaming & Near-Real-Time (Weeks 15–22)

Week 15 — Kafka Fundamentals

- Topics: topics, partitions, replication, producer/consumer acks, consumer groups, ordering.
- Deliverable: Local Kafka via Docker; producer that sends events; consumer that logs lag.

Week 16 — Streaming Compute (Spark Structured Streaming or Flink)

- Topics: event time vs processing time, watermarking, windows, exactly-once with checkpoints + idempotent sinks.
- Deliverable: Streaming job: Kafka → Delta silver; handle late events with watermarking.

Week 17 — Enrichment & Joins in Streams

- Topics: stream-stream, stream-static joins; dedup; upsert to sink.
- Deliverable: Clickstream enrichment with a static dimension table; prove correctness with tests.

Meritshot
EDUCATION

Week 18 — Backpressure & Reliability

- Topics: autoscaling, max offsets per trigger, retries with jitter, DLQs, replay.
- Deliverable: Stress test: intentionally slow consumer; show backpressure metrics and recovery plan.

Week 19 — Monitoring & Alerting

- Topics: Prometheus/Grafana, CloudWatch/Stackdriver, logs/metrics/traces, SLOs.
- Deliverable: Dashboards: throughput, lag, error rate, watermark delay; alert rules.

Weeks 20–22 — Streaming Project #2

- Real-time Orders: Kafka → Streaming job → Gold table (Delta) + mini API (FastAPI) to fetch the latest aggregates**.**

- **KPIs:**
  - At-least-once with idempotent sink
  - End-to-end latency P95 < 5–30s (lab)
  - Recovery from failure < 2 minutes
  - Clear DLQ & replay procedure

Milestone C

- A robust streaming pipeline with metrics, alerts, and incident runbook.

Meritshot
EDUCATION

Phase 4 — Platform, Security, and Infra (Weeks 23–28)

Week 23 — Containers & Packaging

- Topics: Dockerfiles, base images, multi-stage builds, slim images, vulnerabilities.
- Deliverable: Containerize PySpark jobs & Airflow workers; push to registry.

Week 24 — CI/CD

- Topics: GitHub Actions/Azure DevOps/GitLab CI, environment promotion, artifacts.
- Deliverable: CI pipeline: lint, unit/data tests, build image, deploy to dev.

Week 25 — Infra as Code

- Topics: Terraform (S3, IAM roles, MSK/EMR, KMS), state mgmt.
- Deliverable: Terraform modules to spin up storage + a small MSK cluster + minimal EMR/Spark.

Meritshot
EDUCATION

Week 26 — Security Basics

- Topics: IAM roles & least privilege, KMS encryption, secret managers, PII minimization, tokenization.
- Deliverable: Threat model + checklist; encrypt at rest + in transit; secret rotation demo.

Week 27 — Cost & FinOps

- Topics: tagging, budgets, cost alerts, right-sizing, spot, autoscaling.
- Deliverable: Cost dashboard with tag-based allocation (per project/component).

Week 28 — Observability Deep Dive

- Topics: OpenTelemetry basics, structured logs, correlation IDs; data lineage (OpenLineage/Marquez).
- Deliverable: Lineage graph from Airflow → tables; trace a run end-to-end.

Milestone D

- Platform that looks “three-years mature”: containers, CI/CD, IaC, IAM, cost guardrails, observability.

Meritshot
EDUCATION

Phase 5 — Analytics Engineering & Serving (Weeks 29–31)

- dbt models for business metrics (orders, revenue, retention), tests & documentation.
- Dimensional models layered over gold tables; snapshots for SCD.
- Deliverable: dbt project with CI (tests on PR); BI dashboard wired to dbt models.

Phase 6 — Polishing & Interviews (Weeks 32–36)

Week 32 — System Design for Data

- Topics: batch vs streaming, lake vs warehouse vs lakehouse, CDC, global scale, multi-AZ/region, CAP/PACELC, idempotency.
- Deliverable: 2 one-pagers: (1) Real-time analytics design, (2) Backfill strategy for a late-arriving data disaster.

Meritshot
EDUCATION

Week 33 — PySpark/SQL Grind

- 100 targeted questions: joins, windows, performance gotchas; re-implement your hardest transformations 3 ways and compare plans.

Week 34 — Behavioral + Portfolio

- STAR stories: reliability incident, cost reduction, schema evolution conflict, cross-team deliverables.
- Update resume: impact bullets with metrics (latency ↓, cost ↓, SLA ↑). Link repos + READMEs.

Week 35 — Mock Interviews (2–3 rounds)

- One coding (SQL + PySpark), one system design, one behavioral. Capture every gap—fix within 72 hours.

Week 36 — Applications & Referrals

- Target roles listing 2–3 years DE; tailor resume per JD, add a skills matrix; reach out for referrals with a 4-line value message + your repo links.

Meritshot
EDUCATION

Milestone E (Final)

- 2 finished projects (batch + streaming) + platform glue, with metrics, cost notes, and runbooks.
- Comfortably answer “how would you scale / reduce cost / recover from failure”.

Fast-Track (24 Weeks, 12–15 hrs/week)

- Merge phases by doubling weekend build time:
  - Weeks 1–4: SQL+PySpark (compressed)
  - Weeks 5–8: Airflow + Delta + Incremental + DQ
  - Weeks 9–11: Batch Project #1
  - Weeks 12–14: Kafka + Streaming + Observability
  - Weeks 15–17: Streaming Project #2
  - Weeks 18–20: Containers, CI/CD, IaC, Security
  - Weeks 21–22: dbt + BI
  - Weeks 23–24: Design prep, mocks, applications

Meritshot
EDUCATION

Project Portfolio (what recruiters want to see)

Retail Batch Analytics (Airflow + Delta + DQ + dbt)

- Readme: architecture, lineage, SLAs, costs (before/after OPTIMIZE), P95 job times, failure simulation results.

Real-time Orders (Kafka → Structured Streaming → Delta + API)

- Readme: exactly-once strategy, watermark settings, lag dashboard, DLQ replay demo.

Platform Glue (CI/CD + Terraform + IAM + Cost)

- Readme: infra modules, environments, deployment pipeline, permissions, budget alerts.

Interview Drills (use these talk tracks)

- Optimization story: “Shuffle blew up because of key skew; added salting + broadcast hash join; shuffle read ↓ 72%, P95 ↓ 41%.”

- Reliability story: “Backfill created duplicates; added idempotent keys + dedupe on composite (order_id, event_ts); wrote a replay runbook.”

- Cost story: “Compaction + Z-ORDER on customer_id reduced query scan bytes by 65%, saving ~$X/mo.”

- Design trade-off: “Batch for reprocessing & cost; streaming for latency. Same Delta contract so consumers don’t change.”

Meritshot
EDUCATION

Your one-step destination for your Career Upskilling

Lets make a community of 20k+ learnerrs

meritshoteducation