DAY 1 — Foundations of Analytics Engineering & dbt (8 Hours)

Module 1: Introduction to Analytics Engineering (1 hour)

- Evolution of Big Data → Modern Data Stack
- ETL vs ELT (Why ELT dominates Big Data)
- Role of dbt in the Big Data ecosystem
- How dbt integrates with data warehouses/lakehouses (Snowflake, Redshift, BigQuery, Databricks, Fabric)

Module 2: Installing & Setting Up dbt

- dbt Core vs dbt Cloud (choose best for Big Data)
- Installing dbt Core on local machine
- Folder layout & project structure
- Understanding profiles.yml

Module 3: Data Warehouses for Big Data Analytics

- How dbt interacts with Big Data storage
- Connecting dbt to:
    - Databricks
    - Redshift
- Big Data storage: S3, ADLS

Module 4: dbt Basics – Models, SQL, and Materializations

- What are dbt Models?
- SQL-based transformation best practices
- Materializations: table, view, ephemeral, incremental
- When to choose which materialization in Big Data pipelines
- DAGs & dependency graphs

Hands-on:

- Creating basic models
- Running dbt run and viewing DAG

Module 5: Jinja + Macros for Big Data SQL Automation

- Why Jinja matters in analytic pipelines

- Using variables, if conditions, loops
- Creating reusable SQL blocks

Module 6: Source Freshness & Big Data Monitoring

- Declarative source definitions
- Freshness tests
- Source YAML structure
- Integrating with upstream tools like Airflow / ADF / Glue Workflows

Build Your First dbt Project for Big Data:

- Connect to cloud warehouse
- Create Staging Layer
- Create Core Models
- Run & test transformations

DAY 2 — Intermediate to Advanced dbt Modeling for Big Data

Module 7: Staging, Intermediate & Mart Layer Design

- Bronze → Silver → Gold modeling in Big Data
- Naming conventions
- Folder structure for scalable projects
- Reusable staging models

Module 8: Advanced Materializations & Incremental Models

- Understanding partitioning & clustering in Big Data warehouses
- Strategies:
    - Merge incremental
    - Insert incremental
    - Snapshot incremental
- When to use unique_key in big datasets

Hands-on:
Create an incremental model

Module 9: dbt Tests for Data Quality at Scale

- Generic vs Singular tests
- Custom tests in Jinja
- Data quality for Big Data pipelines
- Error handling & pipeline observability

Hands-on:
Test

Module 10: Macros, Packages & Reusable Code

- Creating custom macros
- Big Data macros (date spine, SCD2, dedupe)
- Using popular packages:
    - dbt_utils
    - dbt_expectations
    - dbt_date

Module 11: Documentation & Lineage (1 hour)

- Auto-documentation
- Docs website
- Lineage graphs
- How docsite helps with Big Data governance
- Integrating with Purview

End-of-Day Lab (1 hour)

Build Silver + Gold Layer with Incremental Models

- 3 Layer Medallion Architecture
- Incremental model
- Tests + documentation
- End-to-end ELT run

DAY 3 — Enterprise dbt for Big Data: CI/CD, Governance & Performance

Module 12: Performance Optimization for Big Data

- Optimizing SQL transformations
- Warehouse-specific tuning (Snowflake/BigQuery/Redshift/Databricks)
- Reducing cost in Big Data systems

- Managing compute & storage

Module 13: dbt Snapshots & Slowly Changing Dimensions

- SCD Type 1 & 2
- When to use snapshots in Big Data
- Creating snapshot YAML + SQL
- Performance impact in large datasets

Hands-on:
Snapshot

Module 14: CI/CD for dbt in Big Data Ecosystem

- Using Git + branching strategy
- dbt Cloud jobs
- CI pipelines with:
    - GitHub Actions
    - GitLab CI
    - Azure DevOps
- Automated testing
- Running dbt in production (jobs, schedules)

Module 15: Orchestration with Airflow, ADF, Glue, Databricks

- When to schedule with dbt Cloud vs external orchestrators
- Triggering dbt via:
    - Apache Airflow
    - Azure Data Factory
    - AWS Glue
    - Databricks Workflows

Module 16: Governance, Security & Versioning in Big Data

- Data lineage
- Access control and roles
- Environment management (dev, stage, prod)
- Where dbt fits in end-to-end Big Data governance

Module 17: Project

End-to-End Big Data & dbt Pipeline: