Databricks Tutorial Roadmap

Part 1: The Data Engineering Foundation (Core Skills)

Module 1: Getting Started with the Databricks Workspace

1. How to Launch Your First Cluster and Understand Its Configuration Options.
2. How to Create a Notebook and Use Magic Commands to Mix Python and SQL.
3. How to Navigate the Databricks File System (DBFS) and Upload Your First Dataset.
4. How to Install a Python Library (like pandas or plotly) on a Cluster.
5. How to Clone a GitHub Repository and Manage Notebooks with Git Integration.

Module 2: Mastering Delta Lake

6. How to Convert a Parquet or CSV Dataset into an Optimized Delta Table.
7. How to Perform an *Upsert* Operation using the **MERGE** Command.
8. How to Use Time Travel to Query a Table's State from 24 Hours Ago.
9. How to Recover from a Bad Data Write by Restoring a Table to a Specific Version.
10. How to Add a New Column to a Delta Table Without Breaking Your Pipeline using Schema Evolution.
11. How to Speed Up Queries by Applying **OPTIMIZE** and **Z-ORDER** to a Large Delta Table.
12. How to Clean Up Old Data Versions and Files using the **VACUUM** Command.

Module 3: Building Robust ETL/ELT Pipelines

13. How to Ingest Streaming JSON Data Incrementally Using Auto Loader.
14. How to Configure Auto Loader to Infer and Evolve Schemas Automatically.
15. How to Build a Simple ETL Job Using the Spark DataFrame API.
16. How to Refactor a DataFrame Transformation into a More Readable Spark SQL Query.
17. How to Handle Complex Nested JSON by Flattening it into a Structured Table.
18. How to Create User-Defined Functions (UDFs) to Apply Custom Python Logic at Scale.

Module 4: Working with Streaming Data

19. How to Build a Real-Time Pipeline from a Streaming Source (Kafka, Event Hubs, etc.).
20. How to Aggregate Data in a Stream Using Tumbling Windows.
21. How to Handle Late-Arriving Data Gracefully Using Watermarking.
22. How to Join a Real-Time Stream with a Static Delta Table for Data Enrichment.
23. How to Run a Streaming Job as a One-Time Triggered Batch for Incremental ETL.

Module 5: Declarative Pipelines with Delta Live Tables (DLT)

24. How to Create a Bronze-to-Silver DLT Pipeline that Ingests and Cleans Raw Data.
25. How to Add a Gold Table to Your DLT Pipeline for Business-Level Aggregates.
26. How to Enforce Data Quality Rules in DLT with **@dlt.expect** to Drop or Quarantine Bad Records.
27. How to Visualize the Pipeline Graph and Monitor Data Quality Metrics in the DLT UI.
28. How to Parameterize a DLT Pipeline to Run in Different Environments (Dev vs. Prod).

Module 6: Analytics with Databricks SQL

29. How to Create and Configure a SQL Warehouse for Optimal Query Performance.
30. How to Write and Execute Ad-Hoc Queries in the SQL Editor.
31. How to Build an Interactive Dashboard with Visualizations and Filters.
32. How to Schedule a Dashboard to Refresh Automatically and Email the Results.
33. How to Set Up an Alert to Notify You When a SQL Query Returns a Specific Value.

Part 2: Machine Learning & AI Development

Modules 7–8: Feature Engineering & Model Training

34. How to Train a Scikit-Learn Model on a Single-Node Databricks Cluster.
35. How to Distribute a Simple ML Workload Using `pyspark.ml`.
36. How to Create a Feature Table in the Databricks Feature Store.
37. How to Create a Training Set by Joining Raw Data with Features from the Feature Store.
38. How to Use the Same Feature Store for Both Batch Training and Real-Time Model Serving.

Modules 9–10: Advanced Training and AI Models

39. How to Use AutoML to Generate a Baseline Model and Review the Explanatory Notebooks.
40. How to Distribute Hyperparameter Tuning for an XGBoost Model with Hyperopt and Spark.
41. How to Set Up a GPU Cluster for Deep Learning Workloads.

42. How to Distribute TensorFlow Training Across Multiple Nodes Using Horovod.
43. How to Use Hugging Face Transformers to Perform Inference with a Pre-Trained LLM.
44. How to Build a Basic Retrieval-Augmented Generation (RAG) App by Storing Embeddings in a Delta Table.

Part 3: Production-Grade MLOps

Module 11: End-to-End Tracking with MLflow

45. How to Log Your First Experiment (Parameters, Metrics, and a Model) to MLflow.
46. How to Use the MLflow UI to Compare the Performance of Multiple Training Runs.
47. How to Log a Plot or Image as a Model Artifact for Visual Diagnosis.
48. How to Register Your Best Model to the MLflow Model Registry.
49. How to Transition a Model's Stage from *Staging* to *Production*.
50. How to Add Descriptions and Tags to Your Registered Models for Better Governance.

Module 12: Model Deployment and Serving

51. How to Load the Latest Production Model from the Registry for Batch Scoring.
52. How to Schedule a Daily Job to Score New Data and Write Predictions to a Delta Table.
53. How to Deploy a Model as a Serverless Real-Time Inference Endpoint with One Click.
54. How to Call Your Deployed Model Endpoint from a Python Script Using a REST API Request.
55. How to Interpret the Logs and Performance Metrics of a Model Serving Endpoint.

Modules 13–14: Automation, Governance, and Monitoring

56. How to Securely Store and Access API Keys Using Databricks Secrets.
57. How to Set Up a Basic CI/CD Pipeline with GitHub Actions to Automate Model Training.
58. How to Use Unity Catalog to Grant SELECT Permissions on a Table to Another User.
59. How to Trace a Column's Lineage from a Gold Table Back to its Raw Source using Unity Catalog.
60. How to Build a Simple Dashboard to Monitor for Data Drift in Production.
61. How to Create a Job with Multiple Tasks to Orchestrate a Full ETL + ML Workflow.
62. How to Use the Databricks CLI to Automate Workspace and Job Management from Your Terminal.