SQL SCHOOL — End-to-End Data Engineering with Azure Databricks & Generative AI

SQL • Python • Azure • Databricks • PySpark • GenAI • RAG • Agentic AI

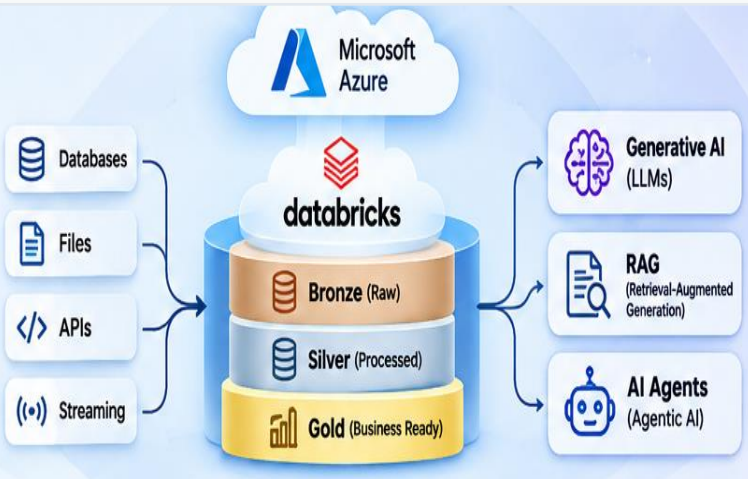
SPECIALIZED CURRICULUM & SYLLABUS

SQL School
Learn . Build . Deploy . Grow.

Azure Data Engineering with Generative AI

SQL | Python | Azure | Databricks | PySpark | GenAI | RAG | AI Agents

A comprehensive, hands-on training program to build end-to-end data engineering and AI-powered data solutions using Azure, Databricks and Generative AI.



What You Will Learn

- Write SQL queries and develop Python applications.
- Build ETL pipelines using SQL, Python and PySpark.
- Work with Azure cloud services and Databricks.
- Build Bronze, Silver and Gold data pipelines.
- Work with Spark, Delta Lake and Unity Catalog.
- Automate and schedule data workflows.
- Implement CI/CD and performance optimization.
- Use Generative AI and LLMs in data engineering.
- Build RAG applications, embeddings and AI agents.

Course Content

The training program is divided into **FOUR courses:**

1. **SQL for Data Engineering**
2. **Python for Data Engineering**
3. **Azure Databricks**
4. **Generative AI & Agentic AI for Data Engineering**

Practical Industry Projects

Project 1 End-to-End Databricks ETL Pipeline Automation.

Project 2 GenAI-Powered Databricks Workflow Automation.

Industry-Recognized Certifications

Prepare for industry certifications from:

- Databricks Certified Data Engineer Associate.
- Databricks Certified Data Engineer Professional.
- Databricks Certified Generative AI Engineer Associate.
- Microsoft Azure Databricks Data Engineer Associate (DP-750).
- Databricks Certified Associate Developer for Apache Spark.

Training Details

Training Mode
Live Online Instructor-Led Training

Course Structure
Four comprehensive courses

Prerequisites
No prior experience required

Learning Approach
Practical training, hands-on exercises and real-world projects.

BUILD REAL SKILLS. WORK ON REAL PROJECTS. BECOME JOB READY. | **SQL SCHOOL** – Learn. Build. Deploy. Grow.

Course 1: SQL for Data Engineering – Course Curriculum

Module 1 – SQL Fundamentals

- Introduction to Databases
- RDBMS Concepts
- SQL Server Overview
- Database Objects
- SQL Server & SSMS Installations
- Creating Databases & Tables
- Data Types
- Constraints
- Primary & Foreign Keys
- Normal Forms

Module 2 – Data Manipulation

- INSERT
- UPDATE
- DELETE
- TRUNCATE
- SELECT Statement
- DISTINCT
- TOP
- ORDER BY

Module 3 – Filtering & Conditional Logic

- WHERE Clause
- AND, OR & NOT
- BETWEEN
- IN
- LIKE
- IS NULL & IS NOT NULL
- CASE WHEN
- Aliases

Module 4 – Aggregate Functions

- COUNT()
- SUM()
- AVG()
- MIN()
- MAX()
- GROUP BY
- HAVING
- Real-World Business Aggregations

Module 5 – SQL Joins

- INNER JOIN
- LEFT JOIN
- RIGHT JOIN
- FULL OUTER JOIN
- SELF JOIN
- CROSS JOIN

- Multi-Table Joins

Module 6 – Advanced SQL Queries

- Subqueries
- Correlated Subqueries
- Common Table Expressions (CTEs)
- UNION
- UNION ALL
- EXISTS
- IN vs EXISTS
- Merge Statement

Module 7 – Window Functions

- ROW_NUMBER()
- RANK()
- DENSE_RANK()
- LEAD()
- LAG()
- OVER()
- PARTITION BY
- Running Totals

Module 8 – Database Programming

- Views
- Stored Procedures
- User-Defined Functions
- Temporary Tables
- Transactions
- Basic Error Handling

Module 9 – SQL Performance & Optimization

- Indexes (Introduction)
- Query Optimization Basics
- Execution Plan Basics
- Best Practices for Writing Efficient SQL
- SQL Coding Standards

Module 10 – End-to-End SQL ETL Project

- Database Design
- Import Raw Data
- Data Cleaning
- Data Validation
- Business Transformations
- Reporting Queries
- Build ETL Pipeline Using SQL
- Generate Final Reporting Tables

Couse 2: Python for Data Engineering – Course Curriculum

Module 1 – Python Fundamentals

- Introduction to Python
- Environment Setup (VS Code & Jupyter Notebook)
- Variables & Data Types
- Input & Output
- Type Casting
- Operators
- Naming Conventions & Best Practices

Module 2 – Python Collections

- Lists
- Tuples
- Sets
- Dictionaries
- List Comprehensions
- Dictionary Comprehensions
- Collection Methods

Module 3 – Conditional Statements & Loops

- if, else & elif
- Nested Conditions
- for Loop
- while Loop
- break, continue & pass
- range()
- enumerate()
- zip()

Module 4 – Functions

- User-Defined Functions
- Function Parameters
- Return Statements
- Lambda Functions
- map() & filter()
- Variable Scope

Module 5 – Strings & File Handling

- String Operations & Methods
- Reading & Writing Text Files
- Working with CSV Files
- Working with JSON Files
- File Handling Best Practices

Module 6 – Exception Handling & Modules

- try, except & finally
- raise Exception
- Python Modules
- Packages
- pip & External Libraries

Module 7 – NumPy

- Arrays
- Indexing & Slicing
- Mathematical Operations
- Basic Statistical Functions

Module 8 – Pandas

- Series & DataFrames
- Reading CSV, Excel & JSON Files
- Data Selection & Filtering
- Sorting & Aggregation
- GroupBy Operations
- Merge, Join & Concat
- Handling Missing Values
- Duplicate Removal
- Data Cleaning & Transformation
- Exporting Processed Data

Module 9 – API & SQL Integration

- REST API Basics
- Reading JSON from APIs
- Python Database Connectivity
- Reading Data from SQL Server
- Writing Data into SQL Tables

Module 10 – End-to-End ETL Project

- Extract Data from CSV, JSON & APIs
- Data Cleaning & Validation
- Business Transformations
- Load Data into SQL Server
- Generate Final Reporting Dataset

Couse 3: Azure Databricks – Course Curriculum

Module 1 – Introduction to Data Engineering:

- Overview of Data Engineering
- Definition and importance
- Key roles and responsibilities
- Introduction to Data Pipelines
- What is a data pipeline?
- Components of a data pipeline
- Examples of data pipelines in the industry

Module 2 – Azure Cloud Fundamentals for Data Engineers

- Delta Lake and Databricks Delta
- Introduction to Microsoft Azure
- Azure Portal navigation
- Azure Resource Groups
- Azure Storage Accounts
- Azure Data Lake Storage Gen2 (ADLS)
- Azure Blob Storage

- Azure Key Vault
- Azure Identity and Access Management (IAM)
- Managed Identities and Service Principals

Module 3 – Introduction to Databricks

- Overview of Databricks
- What is Databricks?
- Databricks vs traditional data platforms
- Setting Up Databricks Environment
- Creating a Databricks account
- Navigating the Databricks workspace
- Databricks Notebooks
- Creating and managing notebooks
- Using markdown and code cells

Module 4 – Data Ingestions and Transformation with Databricks

- Data Ingestion Techniques
- Reading data from various sources (CSV, JSON, Parquet)
- Connecting to databases
- Data Transformation
- Basic transformations using Databricks
- Using SQL in Databricks
- Handling Missing Data and Duplicates
- Techniques for dealing with missing values
- Removing and handling duplicates

Module 5 – Apache Spark and Pyspark

- Overview of Apache Spark
- What is Apache Spark?
- Spark ecosystem and components
- Introduction to PySpark
- Setting up PySpark in Databricks
- PySpark vs Pandas

Module 6 – Pyspark Basics

- PySpark DataFrames
- Creating DataFrames
- Performing basic operations on DataFrames
- DataFrame Transformations and Actions
- Common transformations (select, filter, groupBy, etc.)
- Actions (collect, show, count, etc.)

Module 7 – Advanced Pyspark Concepts

- Working with Spark SQL
- Using SQL queries in PySpark
- Integrating SQL and DataFrame API
- UserDefined Functions (UDFs)
- Creating and using UDFs
- Performance considerations

Module 9 – Data Aggregation and Analysis with Pyspark

- Aggregation Functions
- Grouping and aggregating data

- Window functions
- Data Joins in PySpark
- Different types of joins (inner, outer, etc.)
- Best practices for joins

Module 10 – Databricks Jobs and Workflows

- Introduction to Databricks Jobs and Workflows
- Creating and configuring Jobs and Tasks
- Multi-task workflows and task dependencies
- Scheduling and triggering jobs
- Job parameters and compute configuration
- Monitoring job runs and troubleshooting failures
- Retries, timeouts and notifications

Module 11 – Advanced Topics in Databricks and Pyspark

- Delta Lake and Databricks Delta
- Introduction to Delta Lake
- Implementing Delta Lake in Databricks
- RealTime Data Processing with Structured Streaming
- Basics of Structured Streaming
- Building and managing streaming pipelines

Module 12 – Delta Lake and Lakehouse Architecture (Advanced)

- Introduction to Data Lakehouse Architecture
- Medallion Architecture (Bronze, Silver and Gold)
- Delta Lake Architecture
- Delta Tables
- ACID Transactions
- Time Travel
- Schema Enforcement
- Schema Evolution
- MERGE, UPDATE and DELETE operations
- Delta Live Tables (now Lakeflow Spark Declarative Pipelines)
- Data Quality Expectations
- Change Data Feed (CDF)

Module 13 – CI/CD and Databricks Deployment

- Introduction to Git and GitHub
- Version Control for Databricks Notebooks
- Databricks Git Folders
- Databricks Asset Bundles
- Development, Testing and Production Environments
- CI/CD using Azure DevOps
- Automated Deployment
- Environment Configuration
- Secrets Management
- Pipeline Deployment Best Practices

Module 14 – Databricks Performance Optimization

- Understanding Spark Execution Plans
- Catalyst Optimizer
- Adaptive Query Execution (AQE)

- Broadcast Joins
- Data Skew Handling
- Small File Problems
- Delta OPTIMIZE
- Z-Ordering and Liquid Clustering
- Photon Engine
- Cluster Configuration and Optimization
- Cost Optimization

Couse 4: Generative AI & Agentic AI for Data Engineering – Course Curriculum

Module 1 – Introduction to Generative AI for Data Engineering

- Introduction to Generative AI
- Understanding Large Language Models (LLMs)
- How Generative AI impacts Data Engineering tasks
- Role of Generative AI in the Data Engineering lifecycle
- When to use and when not to use Generative AI
- Understanding AI-assisted data engineering workflows
- Introduction to Agentic Data Engineering
- Understanding the AI Data Stack

Module 2 – Environment Setup and Data Engineering Fundamentals

- Setting up Python
- Setting up VS Code and Jupyter Lab
- Working with Google Colab
- Setting up OpenAI API
- Introduction to Databricks Free Edition
- Databricks workspace navigation
- Creating schemas and volumes
- Loading sample datasets
- Working with CSV files
- Working with JSON files
- Command-line utilities for file processing
- Filtering CSV and JSON data
- Working with Python for data processing
- Introduction to Pandas
- JSON Schema validation
- Introduction to relational databases
- Introduction to NoSQL databases
- PostgreSQL fundamentals
- Dimensional modelling
- Working with JSON in PostgreSQL

Module 3 – Generative AI for Data Generation and Augmentation

- Introduction to AI-based data generation
- Generating synthetic data
- Augmenting existing datasets
- Generating time-series data
- Generating edge cases for data engineering
- Handling personally identifiable information (PII)
- Balancing imbalanced datasets

- Using Generative AI for data augmentation
- Building a data augmentation application
- Creating backend functions
- Running the backend
- Adding frontend components
- Running the completed web application

Module 4 – Generative AI for Data Engineering Code Development

- Using Generative AI to write data engineering code
- Generating Python code with LLMs
- Generating SQL scripts
- Generating Bash scripts
- Data cleaning with Generative AI
- Data modelling with Generative AI
- Creating data schemas
- Creating data systems and pipelines
- Data loading and transformation
- Data quality assessment using LLMs
- Generating code documentation
- Transferring data with Generative AI
- Using Generative AI to fix and improve code
- Using ChatGPT for Data Engineering
- Building Data Engineering applications with Claude
- Custom GPTs for Data Engineering
- Custom LLM and Generative AI tools
- Gemini for BigQuery

Module 5 – Generative AI for Data Parsing and Extraction

- Introduction to AI-based data parsing
- Parsing structured and unstructured data
- Extracting information from PDFs
- Extracting information from web pages
- Extracting data from images
- Parsing HTML and tables
- OCR and scanned document processing
- Extracting information from contracts
- Extracting information from invoices and receipts
- Named Entity Recognition (NER)
- Cleaning extracted data
- Normalization of extracted information

Module 6 – Generative AI for Data Querying and Analysis

- Querying data with Generative AI
- Generating SQL queries using AI
- Optimizing data queries
- Analysing data with Generative AI
- Developing data engineering query applications
- Running query applications
- Converting query applications into web applications
- Adding frontend components
- Combining Excel files with database data

- Generating research reports
- Drafting and sharing reports
- Scheduling tasks and email delivery

Module 7 – AI-Ready Data Engineering and Embedding Pipelines

- Introduction to AI-ready data engineering
- Understanding embeddings
- Embedding models
- Generating embeddings using code
- Similarity and semantic search
- Data ingestion patterns for retrieval
- Data freshness for AI applications
- Data contracts for embeddings
- Schema evolution
- Parsing PDFs, HTML and tables for AI applications
- OCR for scanned documents
- Document cleaning and normalization
- Fixed-size chunking
- Recursive chunking
- Semantic chunking
- Structural chunking
- Chunk size and overlap
- Metadata and parent-child chunking
- Building configurable chunking pipelines
- Batch embedding jobs
- Incremental embedding jobs
- Idempotent embedding pipelines
- Re-embedding when models change
- Backfilling embeddings
- Embedding pipeline cost control

Module 8 – Vector Stores and Hybrid Retrieval

- Introduction to vector stores
- Working with pgvector
- Introduction to Databricks Vector Search
- Introduction to Pinecone and Qdrant
- Vector indexing fundamentals
- HNSW and IVF indexing
- Choosing a vector store
- Row-level security in vector stores
- Multi-tenancy in vector stores
- Vector search and keyword search
- BM25
- Reciprocal Rank Fusion
- Reranking
- Metadata filtering
- Retrieval quality improvement

Module 9 – Text-to-SQL and Semantic Layers

- Introduction to Text-to-SQL
- Natural language to SQL

- Understanding the semantic gap
- Why generated SQL can produce incorrect business answers
- Metrics and dimensions
- Defining business metrics
- Semantic models
- Semantic models as code
- Governed Text-to-SQL
- Verified queries
- Query guardrails
- Read-only query access
- Row limits and allowed tables
- Row-level security in Text-to-SQL
- Building a Text-to-SQL accuracy evaluation harness
- Introduction to Snowflake Cortex Analyst
- Introduction to Databricks Genie Spaces
- Genie instructions and Unity Catalog grounding
- Building a Genie Space

Module 10 – Databricks Genie, Genie Code and AI/BI

- Introduction to Databricks AI tools
- Genie Code and Genie One
- Understanding Genie Code
- Using Genie Code to explore data
- Generating and executing SQL with Genie Code
- Working with customer reviews
- AI Functions
- Classifying and scoring customer reviews
- Introduction to Genie Agents
- Creating Genie Agents
- Understanding agent ins tructions and descriptions
- Teaching agents business terminology
- Synonyms, joins, measures and filters
- Adding example queries
- Benchmarking Genie Agents
- Analysing benchmark results
- Improving agent accuracy
- Collecting user feedback
- Debugging incorrect answers
- Sharing, cloning and certifying agents
- Converting Genie One chats into agents
- Introduction to Metric Views
- Creating Metric Views
- Defining fields, measures and checks
- Connecting Metric Views to Genie Agents
- Creating AI/BI dashboards
- Publishing and sharing dashboards
- Ask Genie on dashboards
- Creating alerts on governed KPIs
- Scheduling tasks and alerts
- Unity Catalog grants

- Row filters and column masks
- Testing permissions through Genie

Module 11 – RAG and Agentic Data Engineering

- Introduction to RAG
- RAG vs Fine-Tuning vs Agents
- Understanding when to use RAG and when to use agents
- Introduction to agent architecture
- Tools and planning
- Agent loops
- Context engineering
- Token budgets
- Retrieval as context
- Building a context assembler
- Introduction to Model Context Protocol (MCP)
- Understanding MCP servers
- Exposing a warehouse as an MCP tool
- Authentication and scopes
- Safe tool design
- Building a data MCP server
- Introduction to agent memory
- Short-term and long-term memory
- Vector and structured memory stores
- Adding memory to agents
- Multi-source agentic retrieval
- Retrieving data from warehouses, vector stores and APIs

Practical Projects – Databricks Jobs & Workflows Automation

Project 1 – End-to-End Databricks ETL Pipeline Automation

- Ingest raw CSV files into Databricks
- Build Bronze, Silver and Gold layers
- Perform data cleaning and transformations using PySpark
- Store processed data in Delta tables
- Create a multi-task workflow with task dependencies
- Schedule daily pipeline execution
- Monitor job runs and troubleshoot failures

Project 2 – GenAI-Powered Databricks Workflow Automation.

- Generate PySpark code using GenAI
- Build and configure Databricks workflows with AI assistance
- Use GenAI to troubleshoot job failures
- Analyse execution logs and generate error explanations
- Use AI suggestions to improve workflow efficiency

Industry-Recognized Certifications You Can Prepare for with this Courses

Upon completion of the training, students will be prepared to pursue the following certifications:

Databricks Certifications

- **Databricks Certified Data Engineer Associate – Preparation for the official certification exam, covering data ingestion, PySpark, Delta Lake, Lakeflow Jobs, data transformations, monitoring and governance.**
- **Databricks Certified Data Engineer Professional – Advanced certification preparation covering production-grade data engineering, advanced pipeline development, performance optimization, troubleshooting and governance.**
- **Databricks Certified Generative AI Engineer Associate – Preparation for the official certification exam, covering LLM application development, RAG architecture, Vector Search, Model Serving, MLflow, AI evaluation and governance.**
- **Databricks Certified Context Engineer Associate – Preparation covering context engineering for AI agents, retrieval systems, agent memory, MCP integration, context management and governance.**
- **Databricks Certified Associate Developer for Apache Spark – Preparation covering Apache Spark architecture, Spark SQL, PySpark DataFrames, Structured Streaming, troubleshooting and performance tuning.**

Microsoft Certifications

- **Microsoft Certified: Azure Databricks Data Engineer Associate (DP-750) – Preparation covering Azure Databricks environment configuration, data ingestion, pipeline development, Unity Catalog governance, deployment and workload monitoring**

BUILD REAL SKILLS. WORK ON REAL PROJECTS. BECOME JOB READY.

Wish you all the best!