AI Data Engineering Program Syllabus

Self-Directed Certificate Program

Coursera Plus | No DeepLearning.AI Track

Mirrors WGU B.S. Data Analytics Structure

Core Tool Stack

Python | MySQL | Docker | VS Code

11 Courses | 5 Portfolio Projects | 3 Terms + GenAI Layer

Industry-Aligned for AI Data Engineer Roles

TERM 1 -- FOUNDATIONS

1. AI & Analytics Foundations

Provider: IBM

Topics Covered:

- Data lifecycle
- Analytics workflows
- Business framing
- Data-driven decisions

Replaces: AI for Everyone. Conceptual overview -- no hands-on coding required.

2. Databases and SQL for Data Science with Python

Provider: IBM

Topics Covered:

- SQL (DDL, DML, joins)
- Python + SQL integration (Jupyter)
- Data modeling basics
- MySQL + SQLite environments

Uses MySQL as primary database. Labs run in Jupyter Notebooks (compatible with VS Code Jupyter extension).

3. Python for Everybody Specialization

Provider: University of Michigan

Topics Covered:

- Python 3 fundamentals
- Data structures
- Web data access
- Databases with Python (SQLite)
- Capstone: data retrieval & visualization

Recommends VS Code as primary editor. Uses SQLite (concepts transfer to MySQL). 5-course specialization.

Alternative: Python for Data Science, AI & Development (IBM)
Provider: IBM -- Single course, more data-science focused.

PORTFOLIO PROJECT: Business Analytics Dashboard Database

Design and build a normalized relational database (MySQL) for a real-world dataset (e.g., e-commerce sales, public health data). Write Python scripts to ingest, clean, and query the data. Produce a short analytics report with business recommendations.

Deliverables:

- GitHub repo with SQL schema and Python scripts
- ERD diagram
- Written summary report with business recommendations

Skills: SQL DDL/DML, Python-SQL integration, data lifecycle, business framing

TERM 2 -- ANALYSIS & MODELING

4. Applied Data Science with Python Specialization

Provider: University of Michigan

Topics Covered:

- pandas & data wrangling
- Statistics & probability
- Visualization (matplotlib, seaborn)
- scikit-learn introduction
- Text mining (NLTK) & network analysis (NetworkX)

5-course specialization. All labs in Jupyter Notebooks (run locally in VS Code). More academic than IBM equivalent.

5. Applied Machine Learning in Python

Provider: University of Michigan

Topics Covered:

- Regression (linear, ridge, lasso)
- Classification (KNN, SVM, decision trees)
- Clustering (K-means)
- Model evaluation & cross-validation
- Ensemble methods (random forests)

Part of the Applied Data Science specialization. Replaces DeepLearning.AI ML Specialization. Classical ML focus using scikit-learn.

6. Data Visualization & Storytelling

Provider: IBM / UC Davis

Topics Covered:

- matplotlib, seaborn, Folium (maps)
- Plotly & Dash (interactive dashboards)
- Choropleth maps & word clouds

Option A: Data Visualization with Python (IBM) -- Python-based.
Option B: Data Visualization with Tableau Specialization (UC Davis) -- dashboard tool.

PORTFOLIO PROJECT: End-to-End ML Analysis with Visual Storytelling

Select a public dataset (Kaggle, UCI ML repo, or government open data). Perform full EDA with cleaning, wrangling, and statistical summaries. Build and evaluate at least 2 ML models. Create a polished visualization dashboard telling the data story.

Deliverables:

- Jupyter notebook with full analysis pipeline
- Model comparison report
- Interactive or static dashboard
- GitHub repo with README walkthrough

Skills: pandas, scikit-learn, model evaluation, data visualization, storytelling

TERM 3 -- ENGINEERING & DEPLOYMENT

7. ETL and Data Pipelines with Shell, Airflow & Kafka

Provider: IBM

Topics Covered:

- Bash/shell scripting for ETL
- Apache Airflow DAGs
- Apache Kafka (streaming)
- cron job scheduling
- MySQL in Docker containers

Best tool-stack match: MySQL + Docker used directly. Theia Cloud IDE (VS Code-like). High ROI for automation roles.

8. Data Engineering, Big Data, and ML on Google Cloud

Provider: Google Cloud

Topics Covered:

- BigQuery (serverless data warehouse)
- Dataflow (Apache Beam)
- Scalable pipelines
- Cloud storage & architecture

NOTE: Uses BigQuery (not MySQL) and Cloud Shell/Qwiklabs (not VS Code). GCP ecosystem knowledge is valuable for cloud roles. Adapt projects to MySQL/Docker locally.

9. Deployment & MLOps

Provider: Google Cloud

Topics Covered:

- Deployment architecture
- Model monitoring
- Production systems
- Model lifecycle management

Option A: Machine Learning in Production on Google Cloud.
Option B: Preparing for Google Cloud Professional Data Engineer Exam.

Both use GCP tools. Apply concepts locally with Docker + Python.

PORTFOLIO PROJECT: Automated Cloud Data Pipeline with Deployed ML Model

Build an ETL pipeline using Apache Airflow that ingests data from an API or file source on a schedule. Transform and load data into MySQL (Docker). Train an ML model and deploy it as a prediction endpoint using Flask/FastAPI in a Docker container. Add basic monitoring and logging.

Deliverables:

- GitHub repo with Airflow DAGs and transformation scripts
- Docker Compose config for full stack
- Architecture diagram
- Demo video or write-up showing pipeline running end-to-end

Skills: ETL orchestration, Airflow, MySQL, Docker, cloud concepts, MLOps, monitoring

GENAI LAYER

10. Introduction to Generative AI

Provider: Google Cloud

Topics Covered:

- What is generative AI
- How LLMs work
- Responsible AI principles

Short but foundational. Conceptual course -- no tool conflicts.

11. Generative AI for Data Analysts

Provider: IBM

Topics Covered:

- AI-powered data analysis workflows
- Natural language to SQL
- Automated EDA and reporting
- Synthetic data generation

More applied and practical for data pipelines. Uses IBM AI tools (watsonx, ChatCSV). Apply concepts with Python + OpenAI/local LLM APIs.

Alternative: ChatGPT Advanced Data Analysis (Vanderbilt University)
Provider: Vanderbilt -- More production-oriented prompt engineering focus.

PORTFOLIO PROJECT: GenAI-Augmented Data Analysis Tool

Build a Python application that uses an LLM API (OpenAI, Google Gemini, or local model) to assist with data analysis. Implement prompt-engineered workflows: natural language to SQL (MySQL), automated EDA summaries, and report generation. Integrate into a dataset or pipeline from a previous project.

Deliverables:

- GitHub repo with working application
- Example prompts and outputs documented
- Write-up comparing GenAI-assisted vs. manual analysis

Skills: LLM API integration, prompt engineering, practical GenAI, pipeline augmentation

CAPSTONE

PORTFOLIO PROJECT: Full-Stack AI Data Engineering Solution

Combine all skills into one end-to-end project: Ingest real-world data from multiple sources (API + MySQL database + files). Build a scheduled ETL pipeline (Airflow in Docker). Apply ML models for prediction or classification. Deploy the model with monitoring (Docker + Flask/FastAPI). Add a GenAI layer (natural language querying or automated reporting). Present results in a dashboard.

Deliverables:

- GitHub repo with complete source code
- Architecture diagram
- Docker Compose for full local deployment
- Live demo or recorded walkthrough
- Portfolio-ready technical write-up

Skills: SQL, Python, ML, ETL, Docker, Airflow, MLOps, GenAI, data storytelling

COMPLETE COURSE LIST

#	Course	Provider
1	Introduction to Data Analytics	IBM
2	Databases and SQL for Data Science with Python	IBM
3	Python for Everybody Specialization	U of Michigan
4	Applied Data Science with Python	U of Michigan
5	Applied Machine Learning in Python	U of Michigan
6	Data Visualization with Python	IBM
7	ETL & Data Pipelines w/ Shell, Airflow & Kafka	IBM
8	Data Engineering, Big Data, and ML on GCP	Google Cloud
9	Machine Learning in Production on GCP	Google Cloud
10	Introduction to Generative AI	Google Cloud
11	Generative AI for Data Analysts	IBM

PORTFOLIO SUMMARY

Project	Term	Key Skills
Business Analytics Dashboard Database	Term 1	SQL, Python, ERD, business analysis
End-to-End ML Analysis	Term 2	pandas, scikit-learn, visualization, EDA
Automated Cloud Data Pipeline	Term 3	Airflow, MySQL, Docker, MLOps
GenAI-Augmented Analysis Tool	GenAI	LLM APIs, prompt engineering, automation
Full-Stack AI Data Engineering	Capstone	All skills combined

GRADUATE COMPETENCIES

Skill	Source
SQL + Schema Design	Courses 2, 4
Python Automation	Courses 3, 4
Classical Machine Learning	Course 5
Cloud Data Engineering	Course 8
ETL Orchestration (Airflow)	Course 7
Production ML Architecture	Course 9
GenAI Workflow Integration	Courses 10, 11

Program Notes

- Mirrors WGU B.S. Data Analytics structure and progression
- Avoids DeepLearning.AI dependency
- Fully Coursera Plus compatible
- Industry-aligned for AI Data Engineer roles
- 5 portfolio projects demonstrating progressive skill building
- Core stack: Python, MySQL, Docker, VS Code
- GCP courses teach cloud concepts -- adapt hands-on work to local Docker/MySQL stack