Data Engineering with Python (Capstone-Aligned)

Introduction

This 18-day Data Engineering with Python course blends theory with hands-on practice. Students learn core concepts in Python, data analysis, statistics, machine learning, and deployment while progressively building a real-world **Result Data Analytics Platform** as their capstone project.

Objectives

- Strengthen foundations in Python, statistics, databases, and ML.
- Apply daily learning through concept-focused and capstone-linked assignments.
- Develop collaborative coding, version control, and deployment skills.
- Deliver a complete end-to-end data science solution by course end.

How to Achieve It

- **Dual Assignments:** Daily tasks reinforce concepts and build capstone components.
- **Stepwise Integration:** Each artifact (scripts, queries, models, dashboards) adds to the final project.
- **Tools & Practices:** Use Python, SQL/MongoDB, Scikit-learn, TensorFlow/PyTorch, Streamlit, and GitHub for a professional workflow.

Day 1 – 25-09-2025

Topic: Python Basics & Data Structures

- Python functions and modular design for reusability.
- Data structures: lists, dicts, sets, tuples for student records.
- Comprehensions, iterators, and generators for concise coding.
- **Assignment-1:** Implement student record operations (CRUD) using nested dictionaries.
- **Assignment-2:** Create students_sample.csv and data_loader.py to load/save structured student data (name, seat number, subject, marks).

Day 2 – 26-09-2025

Topic: Data Manipulation & Visualization (Part 1)

- Handling missing values, duplicates, outliers with Pandas.
- Descriptive stats and correlation heatmaps.

- Visualizing distributions with Seaborn.
- **Assignment-1:** Perform EDA on a dataset (e.g., stock or COVID).
- **Assignment-2:** Enhance data_loader.py to clean student data and save students_clean.csv. Add validation for nulls and duplicates.

Day 3 – 27-09-2025

Topic: Data Manipulation & Visualization (Part 2)

- Pivot tables and multi-indexing.
- Interactive visualization with Plotly.
- Feature engineering: deriving new metrics.
- **Assignment-1:** Build interactive Plotly dashboard for sample dataset.
- **Assignment-2:** Create eda_report.ipynb generating plots of student marks distribution, subject averages, and correlation heatmap.

Day 4 – 09-10-2025

Topic: Statistics & Probability (Part 1)

- Distributions (normal, binomial, Poisson).
- Sampling and hypothesis formulation.
- Measures of central tendency and variability.
- **Assignment-1:** Simulate dice/coin experiments and validate distributions.
- **Assignment-2:** Add statistical summaries (summary_stats.py) to compute class mean, variance, subject-wise distributions from student dataset.

Day 5 – 10-10-2025

Topic: Statistics & Probability (Part 2)

- Hypothesis testing (t-test, ANOVA).
- Confidence intervals, p-values.
- Correlation vs causation.
- **Assignment-1:** Run t-test on two groups' exam scores.
- **Assignment-2:** Extend summary_stats.py to test if two classes differ significantly in average marks (ANOVA). Save results in capstone/reports/stats_report.md.

Day 6 – 11-10-2025

Topic: Version Control

- Git basics: commits, branching, merging.
- Conflict resolution & collaboration.
- GitHub workflows & pull requests.
- **Assignment-1:** Create GitHub repo, push daily assignments, simulate conflict resolution.
- **Assignment-2:** Initialize capstone-project repo with folder structure: /data, /scripts, /notebooks, /reports. Commit all artifacts so far.

Day 7 – 23-10-2025

Topic: SQL & Data Extraction

- SQL joins, aggregations, indexing.
- Subqueries & optimization.
- Connecting SQL with Python.
- **Assignment-1:** Practice SQL queries on sample employee dataset.
- **Assignment-2:** Create etl_to_db.py that loads students_clean.csv into students, subjects, and marks tables (MySQL or Mongo). Provide schema file.

Day 8 – 24-10-2025

Topic: Machine Learning Fundamentals (Part 1)

- Preprocessing: scaling, encoding.
- Train-test split, cross-validation.
- Linear & Logistic Regression.
- **Assignment-1:** Logistic regression on Titanic dataset.
- **Assignment-2:** Implement predict_pass_fail.py that predicts student pass/fail from marks. Store predictions in DB.

Day 9 – 25-10-2025

Topic: Machine Learning Fundamentals (Part 2)

- Decision Trees & Random Forests.
- Bias-variance trade-off.

- Model interpretability.
- **Assignment-1:** Train Random Forest on Titanic dataset, report accuracy.
- **Assignment-2:** Apply Random Forest to student dataset, generate feature importance. Save results in capstone/reports/model_report.md.

Day 10 – 30-10-2025

Topic: Advanced AI & Deployment (Part 1)

- Neural networks: perceptron, activations.
- Keras/PyTorch basics.
- Loss functions, optimizers.
- **Assignment-1:** Train a neural net on MNIST digits.
- **Assignment-2:** Create a simple neural net for student grade prediction. Store trained model as capstone/models/grade_nn.h5.

Day 11 – 31-10-2025

Topic: Advanced AI & Deployment (Part 2)

- CNNs for image classification.
- Pooling, dropout, transfer learning.
- **Assignment-1:** Fine-tune ResNet on CIFAR-10.
- **Assignment-2:** Build a capstone/notebooks/student_performance_nn.ipynb comparing ML vs NN models for pass/fail.

Day 12 – 01-11-2025

Topic: Advanced AI & Deployment (Part 3)

- RNNs & LSTMs for sequences.
- Sentiment analysis, embeddings.
- **Assignment-1:** IMDB sentiment classification using LSTM.
- **Assignment-2:** Build an LSTM model predicting student's next subject score given past performance. Save as capstone/models/lstm_model.h5.

Day 13 – 06-11-2025

Topic: Advanced AI & Deployment (Part 4)

- Model deployment with Flask/FastAPI.
- Building REST APIs.
- Containerization with Docker.
- **Assignment-1:** Deploy a ML model as API locally.
- **Assignment-2:** Implement /predict endpoint in app.py serving student pass/fail predictions.

Day 14 – 07-11-2025

Topic: Advanced AI & Deployment (Part 5)

- Streamlit/Gradio apps.
- End-to-end ML pipelines.
- CI/CD workflows.
- **Assignment-1:** Build a Streamlit app for classification.
- **Assignment-2:** Create dashboard.py Streamlit app showing student performance metrics + prediction tool.

Day 15 – 13-11-2025

Topic: Advanced AI & Deployment (Part 6)

- Hyperparameter tuning.
- Handling imbalanced data.
- Explainability with SHAP/LIME.
- Fairness, bias reduction, and responsible AI practices.
- **Assignment-1:** Tune Random Forest with GridSearchCV.
- **Assignment-2:** Apply SHAP on student prediction model, evaluate fairness metrics (e.g., group performance disparities), and generate capstone/reports/explainability.md.

Day 16 – 14-11-2025

Topic: Advanced AI & Deployment (Part 7)

- Clustering methods (K-Means, Hierarchical).
- PCA & dimensionality reduction.
- Anomaly detection.

- **Assignment-1:** Cluster customer segmentation dataset.
- **Assignment-2:** Apply clustering on student marks to group performance bands (high, medium, low). Store in DB.

Day 17 – 15-11-2025

Topic: Advanced AI & Deployment (Part 8)

- Reinforcement learning fundamentals.
- Q-learning algorithm.
- Applications in games, robotics.
- **Assignment-1:** Implement Q-learning on GridWorld.
- **Assignment-2:** Add bonus analytics to dashboard: flagging students needing "intervention" (low score trajectory) inspired by RL decision-making logic.

Day 18 – 20-11-2025

Topic: PDF Data Extraction & Automation

- Extracting text/tables with PyPDF2/pdfplumber.
- Converting PDF data to structured DB.
- Workflow automation for batch processing.
- **Assignment-1:** Extract tabular data from sample PDF.
- **Assignment-2:** Implement pdf_ingest.py pipeline that extracts VTU-like result PDFs → DB → dashboard refresh. Final integration of all modules.

Capstone Project Outline: Student Result Analytics System

1. Project Title: "Result Data Analytics Platform using Database & Visualization"

2. Problem Statement

Universities publish exam results in PDF/mark-sheet formats, which are difficult to analyze for large batches (100–500 students). The challenge is to convert these unstructured results into a structured database and build an application that provides meaningful insights.

3. Objectives

- Convert raw mark-sheet data (PDF/Excel) into a database (MySQL or MongoDB).
- Build a front-end dashboard for data visualization and analytics.
- Provide insights such as:
    - Top 5 and Bottom 5 students overall.
    - Subject-wise toppers and failures.
    - Student-wise detailed performance.
    - Pass/Fail statistics.
    - Aggregate metrics (average, highest, lowest).

4. System Architecture

- **Data Ingestion Layer:** Extract and clean results from PDF/CSV.
- **Database Layer:** Store structured results in MySQL (relational) or MongoDB (document-based).
- **Back-end API Layer:** Node.js / Python (Flask/Django/FastAPI) for queries.
- **Front-end Layer:** React/Angular/Vue with charts (Chart.js, Recharts, D3.js).
- **Visualization:** Dashboard showing insights with filters (student-wise, subject-wise, semester-wise).

5. Modules

1. **Data Import & Cleaning** – Convert PDF mark sheets into database entries.
2. **Student Management** – Store student details, marks, and results.
3. **Analytics Engine** – Compute insights (rankings, pass/fail, averages).
4. **Dashboard UI** – Interactive charts, tables, and reports.

5. **Search & Filters** – Search by student ID, name, or subject.

6. Expected Outcomes

- Functional web application with a connected database.
- Dashboard showing performance insights.
- Automated report generation (PDF/Excel) for analytics.
- Improved understanding of databases, APIs, and visualization tools.

7. Tech Stack (Suggested)

- **Database:** MySQL / MongoDB
- **Back-end:** Python (Flask/Django/FastAPI) / Node.js
- **Front-end:** React.js / Angular
- **Visualization:** Chart.js, Recharts, or D3.js
- **Deployment:** Docker / Cloud