Roadmap for Learning Python as a Data Engineer

1. Python Fundamentals

Learn the basics of Python, including variables, data types, loops, and functions. These are essential for writing scripts and automating tasks.

2. Data Manipulation with Pandas

Use Pandas for data cleaning, transformation, and analysis. Master operations on DataFrames to handle complex datasets.

3. Working with Databases

Learn to connect to databases, execute SQL queries, and use ORMs like SQLAlchemy. This is critical for managing structured data.

4. File Handling

Understand how to read and write files in various formats like CSV, JSON, and Excel.

5. Data Pipelines and Automation

Learn to build and manage ETL pipelines using tools like Airflow and Prefect. Automate repetitive workflows to save time.

6. Working with APIs

Consume REST APIs, handle authentication, and process JSON responses to integrate with external services.

7. Big Data with PySpark

Master PySpark for processing large-scale datasets efficiently.

8. Regular Expressions

Extract and clean data from unstructured formats like text and logs using regular expressions.

9. Data Serialization

Learn formats like JSON, Pickle, and YAML to store and transfer data efficiently.

10. Parallelism and Concurrency

Optimize code performance using multithreading, multiprocessing, and async programming.

11. Cloud Integration

Integrate with cloud services like AWS, GCP, or Azure to manage and store data.

12. Logging and Monitoring

Implement logging and error handling to debug and monitor pipeline performance.

13. Version Control and Collaboration

Use Git for version control to collaborate efficiently in teams.

14. Data Visualization (Optional)

Create visualizations using Matplotlib, Seaborn, or Plotly to explore and present data.