Syllabus

Data engineering on python

Topics covered as follows

Introduction to Data engineering

- Overview of data engineering
- Importance of data engineering in data data ecosystem
- Role of Data engineers in the organization

Python fundamentals for data engineering

- Python basics: syntax ,variables, Data types
- Control flow and loops
- Function and modules
- Exception handling

Working with data in python

- Data manipulation with pandas
- Data cleaning and preprocessing
- Handling missing data
- Exploratory data analysis (EDA)

Introduction to Database

- Relational database overview
- SQL basics for data engineers
- Connecting python to database (sQlite , Mysql)

Building data pipelines with apache airflow

- Introduction to apache airflow
- DAGS (Directed Acyclic Graphs)
- Operators and sensors in the airflow
- Scheduling and monitoring data pipelines

Big data technologies with python

- Introduction to big data and hadoop
- Introduction big data with pyspark
- Working with data frames in pyspark

Data serialization and formats

- Introduction to data serialization
- Jason and xml serialization
- Csv and parquet format

Real world project building an end to end pipeline

- Understanding project requirements
- Designing data pipeline architectures
- Implementing and testing the data pipeline
- Troubleshooting and optimization

Version control and collaboration in data engineering

- Introduction to version control (GIT)
- Collaborative development practice

Best practices and optimization techniques

- Code optimization strategies
- Scalability and performance tuning
- Error handling and logging