Course Title: Data Engineering Essentials

Course Description:

This course focuses on the principles and practices of data engineering, including the management, manipulation, and transformation of data at scale. Students will learn about data modeling, data warehousing, big data technologies, and real-time data processing systems.

Course Objectives:

- Understand the fundamentals of data engineering and big data.
- Learn to model, store, and retrieve data effectively.
- Gain proficiency with big data technologies like Hadoop and Spark.
- Develop skills in building and managing data pipelines.
- Understand best practices for data security and governance.

Prerequisites:

- Basic knowledge of programming (Python preferred).
- Familiarity with SQL and database concepts.
- Understanding of fundamental data structures.

Weekly Syllabus Outline:

Week 1: Introduction to Data Engineering

- Overview of data engineering and its importance.
- Key concepts: Data lakes, data warehouses, ETL processes.
- Setting up the development environment.

Week 2: Data Modeling Concepts

- Principles of data normalization and denormalization.
- Introduction to dimensional modeling.
- Designing schemas for analytical processing.

Week 3: SQL Deep Dive

- Advanced SQL techniques for data manipulation.
- Window functions and complex joins.

- Performance optimization in SQL queries.

Week 4: Introduction to NoSQL Databases

- Overview of NoSQL vs. SQL databases.
- Types of NoSQL databases: Document, key-value, graph, and columnar.
- Use cases and performance considerations.

Week 5: Data Integration (ETL)

- Understanding ETL (Extract, Transform, Load) processes.
- Designing and implementing ETL pipelines.
- Tools and frameworks for ETL (e.g., Apache NiFi, Talend).

Week 6: Data Storage and Retrieval

- Data warehousing solutions (Amazon Redshift, Google BigQuery).
- Building and querying data warehouses.
- Data indexing and partitioning strategies.

Week 7: Big Data Technologies

- Introduction to the Hadoop ecosystem.
- HDFS for data storage, MapReduce for processing.
- Basics of Apache Hive and Pig for data querying and manipulation.

Week 8: Real-Time Data Processing

- Stream processing vs. batch processing.
- Introduction to Apache Kafka for data ingestion.
- Using Apache Storm and Apache Flink for stream processing.

Week 9: Advanced Analytics and Machine Learning

- Integrating data pipelines with machine learning models.
- Overview of machine learning algorithms for big data.
- Using Spark MLlib for predictive analytics.

Week 10: Data Pipeline Orchestration

- Workflow management with Apache Airflow.
- Scheduling and monitoring data jobs.
- Best practices in data pipeline design.

Week 11: Data Security and Governance

- Data privacy and protection laws (GDPR, HIPAA).
- Techniques for data encryption and secure data transfer.
- Data governance frameworks and best practices.

Week 12: Cloud Data Solutions

- Leveraging cloud platforms for data engineering (AWS, Azure, GCP).
- Cloud-specific data services and optimizations.
- Cost management and scalability in the cloud.

Week 13: Advanced Data Engineering Topics

- Introduction to data lakes and their architecture.
- Building a data lake using technologies like Amazon S3 and Azure Data Lake.
- Data cataloging and metadata management.

Week 14: Capstone Project

- Project initiation: Planning and designing a data solution.
- Implementation phase: Building a comprehensive data pipeline.
- Integrating analytics and business intelligence tools.

Week 15: Project Presentations and Course Wrap-Up

- Presentation of capstone projects.
- Peer reviews and feedback sessions.
- Course summary and final evaluations.

Assessment Methods:

- Weekly hands-on labs and coding assignments.
- Regular quizzes to reinforce and test knowledge.
- A comprehensive capstone project that involves creating an end-to-end data engineering solution.