Big Data Analytics

System and Information technology

In-Class

40 hours

EGP 15,000

Course Description:

This course provides a comprehensive introduction to Big Data concepts and practical implementation using Apache Hadoop and Apache Spark. Participants will explore the sources, characteristics, and importance of Big Data, along with the technologies used for its storage and processing. The course delves into Apache Spark's ecosystem, emphasizing Resilient Distributed Datasets (RDDs) and their operations. A crash course in Python programming is included, focusing on essential packages like NumPy, Matplotlib, and Pandas. Through practical labs and real-world projects using PySpark, students will gain hands-on experience in Big Data analytics tasks such as churn prediction and word count analysis.

Target Audience:

- Data Professionals
- IT and Development Personnel
- Python Programmers
- Students and Researchers

Course Objectives:

- Understand the definition, characteristics, and importance of Big Data.
- Learn about Hadoop Distributed File System (HDFS) and MapReduce for Big Data processing.
- Gain proficiency in Apache Spark and its ecosystem components, including SparkSession, RDDs, transformations, and actions.
- Develop foundational Python programming skills with a focus on data manipulation and visualization libraries (NumPy, Pandas, Matplotlib).
- Implement Big Data projects using PySpark, integrating Spark's RDD operations with Python for scalable data processing.

Course Outline:

- Introduction to Big Data & Apache Hadoop
- Big Data tools and techniques
- Introduction to Python Programming
- Python libraries
- Practical Implementation of Big Data using PySpark

Assessment and Attendance Strategy:

- Participants will be evaluated based on their participation in class discussions and individual exercises.
- Each Participant must achieve 80% attendance of the total in-class sessions.

Course language:

Bilingual

Prerequisites:

No prerequisite for this course