PCAC2009 BIG DATA INTEGRATION AND MANAGEMENT (3-0-0)

OVERALL COURSE OBJECTIVES: The overall course objective is to enable learners to effectively understand and handle big data issues, develop skillsets for processing and integrating big data on Hadoop and Spark platforms, and apply machine learning techniques to construct data-driven models and facilitate decision-making process.

LEARNING OUTCOMES: On successful completion of the course the students shall be able to:

1. Ability to recognize different data elements in various scenarios and explain the necessity for a Big Data Infrastructure Plan and Information System Design.
2. Ability to select suitable data models for specific types of data and apply techniques for handling streaming data.
3. Proficiency in retrieving data from different databases and big data management systems, and identifying when a big data problem needs data integration.
4. Capability to execute basic big data integration and processing on Hadoop and Spark platforms.
5. Ability to design a strategy to leverage data using the steps in the machine learning process and apply machine learning techniques to explore and prepare data for modelling.
6. Proficiency in constructing models that learn from data using open source tools and analyzing big data problems using scalable machine learning algorithms on Spark.

COURSE CONTENT:

Module 1: Big Data Modeling and Management Systems [13 Hours]

This course covers how to collect, store, and organize big data using appropriate management tools. It explores a range of data genres, big data platforms, big data management systems, and analytical tools. Guided, hands-on tutorials provide familiarization with techniques using real-time and semi-structured data examples. Systems and tools covered include AsterixDB, HP Vertica, Impala, Neo4j, Redis, SparkSQL. Key learning outcomes include identifying different data elements, designing a Big Data Infrastructure Plan and Information System, handling streaming data, differentiating between a traditional Database Management System and a Big Data Management System, and designing a big data information system. The course is suitable for those new to data science with completion of the Intro to Big Data recommended. Basic installation skills and virtual machine usage are necessary for hands-on assignments.

Sub-Topic
Designing a Big Data Management System for an Online Game
Introduction to Big Data Modeling and Management
Working With Data Models
Exploring Streaming Sensor Data
DBMS-based and non-DBMS-based Approaches to Big Data

Formative Assessments:
4 graded quizzes and 1 Peer-review assignment.

Module 2: Big Data Integration and Processing [18 Hours]

The course covers the process of identifying, collecting, storing, and organizing big data. It explores various data genres, management tools, big data platforms, management systems, and analytical tools. Through hands-on tutorials, learners will get familiar with real-time and semi-structured data examples. The course discusses various systems and tools including AsterixDB, HP Vertica, Impala, Neo4j, Redis, and SparkSQL. By the end, learners will be able to recognize different data elements, understand why a Big Data Infrastructure Plan is necessary, identify frequent data operations, select suitable data models, handle streaming data, differentiate between traditional and big data management systems, and design a big data information system. It is intended for data science beginners. Prior programming experience is not needed, but the ability to install applications and utilize a virtual machine is essential for hands-on assignments.

Sub-Topic
Big Data Analytics using Spark
Big Data Integration
Learn By Doing: Putting MongoDB and Spark to Work
Processing Big Data
Retrieving Big Data

Formative Assessments:
10 graded quizzes

Module 2: Machine Learning With Big Data [22 Hours]
This course provides an introduction to machine learning techniques used to explore, analyze, and utilize data. It offers insights into various tools and algorithms for creating machine learning models that can learn from data and handle big data problems. After completion, learners will be equipped to devise an approach to leverage data using machine learning processes, apply machine learning techniques for data modeling, recognize the type of machine learning problem to implement suitable techniques, create models with widely available open-source tools, and analyze big data problems using scalable machine learning algorithms on Spark.

Sub-Topic

Data Exploration
Data Preparation
Evaluation of Machine Learning Models
Introduction to Machine Learning with Big Data
Regression, Cluster Analysis, and Association Analysis

Formative Assessments:
11 graded quizzes.

ASSESSMENT:

For summative assessments, Coursera will provide question banks for which exams can be conducted on the Coursera platform or the faculty will create their own assessments.

Note: If a Course or Specialization becomes unavailable prior to the end of the Term, Coursera may replace such Course or Specialization with a reasonable alternative Course or Specialization.