ĐẠI HỌC KHOA HỌC TỰ NHIÊN
TRUNG TÂM TIN HỌC

Địa chỉ: 227 Nguyễn Văn Cừ, Q.5, TP.HCM
Tel: (028) 38351056
Email: tuvan@csc.hcmus.edu.vn
Website: https://csc.edu.vn/

Mới

BIG DATA IN MACHINE LEARNING

(DỮ LIỆU LỚN TRONG MÁY HỌC)

Thời gian học: 5 tuần

Thời lượng: 48 giờ (64 tiết)

Học phí: 6.000.000 đ

1. MỤC TIÊU

- Trang bị cho học viên (HV) những kiến thức nền tảng về đặc điểm và các thành phần của Dữ liệu lớn (Big Data)
- Khám phá tiềm năng to lớn của Dữ liệu lớn và vai trò then chốt của PySpark trong việc khám phá những bí mật bên trong nó
- Nắm vững kỹ thuật xử lý các bộ dữ liệu khổng lồ một cách dễ dàng bằng cách sử dụng các công cụ mạnh mẽ của PySpark như RDD, DataFrame, Streaming…
- Bước vào hành trình Máy học (Machine Learning), tận dụng PySpark để triển khai các thuật toán tiên tiến, chuyển đổi dữ liệu thô thành thông tin hữu ích, đưa ra dự đoán
- Tìm hiểu quá trình Xử lý ngôn ngữ tự nhiên (NLP) với PySpark, cho phép diễn giải và phân tích dữ liệu văn bản
- Trang bị kỹ năng để thiết lập và quản lý cụm Spark, giúp HV sẵn sàng giải quyết các thách thức về dữ liệu lớn trong thế giới thực
- Giúp HV nắm bắt được các công nghệ sử dụng trong Dữ liệu lớn: cách lưu trữ, quản lý, xử lý và phân tích dữ liệu lớn để mang lại giá trị cho doanh nghiệp
- *Là khóa cuối trong chương trình "Data Science and Machine Learning Certificate"*

2. ĐỐI TƯỢNG HỌC

- HV học qua khóa "*Machine Learning with Python"* hoặc có kiến thức tương đương
- Sinh viên các trường Đại học, Cao đẳng
- HV có định hướng sẽ làm việc trong lĩnh vực Machine Learning hoặc Data Science

3. KẾT QUẢ ĐẠT ĐƯỢC

Sau khi hoàn thành khóa học, học viên sẽ đạt được các kỹ năng:

- Nắm vững các đặc điểm và thành phần của Dữ liệu lớn
- Nắm vững các kỹ thuật xử lý và phân tích Dữ liệu lớn
- Làm việc với Spark, Big Data Technology mới nhất

- Dễ dàng thao tác với Dữ liệu lớn sử dụng bộ thư viện của PySpark: PySpark RDD's, PySpark DataFrames, PySpark SQL, PySpark ML, PySpark Streaming, PySpark GraphX…
- Áp dụng Máy học với Dữ liêu lớn, dự đoán xu hướng và ra quyết định
- Giải mã sự phức tạp của ngôn ngữ sử dụng công cụ NLP tiên tiến của PySpark
- Xử lý dữ liệu thời gian thực
- Thiết lập và điều hướng cụm Spark, đảm bảo hiệu suất tốt trong các thách thức về dữ liệu
- Vận dụng các kỹ thuật phân tích dữ liệu lớn để mang lại các số liệu thống kê theo yêu cầu của doanh nghiệp

4. NỘI DUNG KHÓA HỌC

- Tổng quan Dữ liệu lớn (Big data)
  - Giới thiệu, lịch sử Big Data
  - Vs' của Big Data (3Vs', 4Vs', 5Vs', 6Vs'…)
  - So sánh Batch processing (xử lý theo lô) và Stream processing (xử lý theo thời gian thực)
  - Giới thiệu Apache Spark
  - Các thành phần của Apache Spark: RDD API, SQL, MLlib, GraphX, Streaming
- Tổng quan PySpark
  - Giới thiệu PySpark: Spark với Python (Python API)
  - Lý do chọn PySpark
  - Cài đặt và cấu hình PySpark
  - Spark Context, Spark Session
- PySpark RDDs
  - Giới thiệu PySpark RDDs (Resilient Distributed Dataset)
  - RDDs operations
    - Transformations
    - Actions
  - Làm việc với PySpark RDDs
    - RDDs
      - Tạo RDD: parallelize(), textFile()
      - Transformations: map(), filter(), flatMap(), RDD1.union(RDD2)…
      - Actions: collect(), take(), count(), first(), reduce(), saveAsTextFile(),…
    - Pair RDDs
      - Tạo Pair RDDs: từ key-value tuple, regular RDD
      - Transformations: reduceByKey(), groupByKey(), sortByKey(), join()
      - Actions: countByKey(), collectAsMap()
- PySpark DataFrame, SQL
  - Giới thiệu
  - Làm việc với PySpark DataFrame
    - Tạo DataFrame: createDataFrame(), spark.read.csv(), spark.read.json()…
    - Các function thông dụng
      - printSchema(), show(), count(), describe(), crosstab()

        - select(), select() và agg, count, max, mean, min, sum..., select().distinct(),
        - groupby(), orderby().asc()/desc()
        - withColumn(), withColumnRenamed()
        - drop(), dropDuplicates(), dropna()
        - filter(), where()
        - Conditional clauses: .when(<if condition>, <then x>), .otherwise()
        - User defined functions (UDF)
    - Trực quan hóa dữ liệu
  - PySpark SQL
    - Giới thiệu
    - Truy vấn thông dụng: select(), when(), like(), startswith(), endswith(), substr(), between()
    - Thao tác trên dữ liệu: tạo view từ dataframe, nhóm (group by), lọc (filtering), sắp xếp (sorting), kết (joining), phân vùng (partitioning)…
- Tiền xử lý và phân tích dữ liệu
  - Xử lý dữ liệu
    - Xóa (dropping), lọc (filtering), kết (joining) dữ liệu
    - Xử lý dữ liệu thiếu, trùng, outliers
    - Sử dụng parquet
    - Data validation
  - Tạo các tính năng mới (Feature Engineering)
    - Từ tính năng kiểu chuỗi/ số -> các tính năng kiểu số mới
    - Từ tính năng chuỗi -> tính năng Datetime -> các tính năng thành phần từ Datetime
    - Trích xuất tính năng chuỗi/ văn bản thành các tính năng mới
    - Splitting & Exploding
    - Scaling data
    - Pivoting & Joining
    - Binarizing, Bucketing & Encoding
  - Phân tích dữ liệu (Data Analysis)
    - Phân tích dữ liệu khám phá (EDA)
    - Trực quan hóa dữ liệu
- Tổng quan PySpark MLlib
  - Giới thiệu
  - PySpark MLlib algorithms
  - Triển khai project: Đọc và xử lý dữ liệu, xây dựng model, đo lường và đánh giá model, lưu model và dự đoán mới
- Machine Learning với PySpark MLlib
  - Học có giám sát (Supervised Learning: Classification & Regression)
    - Linear Regression, Logistic Regression
    - Tree models: Decision Tree, Random Forest, Gradient-Boosted Tree
  - Pipeline
  - Học không giám sát (Unsupervised Learning: Clustering & Recommender System)
    - Phân cụm với KMeans
    - Hệ thống đề xuất (Recommender System) với ALS
    - Phân tích luật kết hợp (Association rules) với FPGrowth
- Xử lý ngôn ngữ tự nhiên (Natural Language Processing – NLP)

- Giới thiệu
- Công cụ
  - Tokenizer
  - StopWordsRemover
  - NGram
  - CountVectorizer
  - TF-IDF

- Apache Spark standalone cluster
  - Giới thiệu standalone cluster
  - Kết nối các Slave computer tới Master Server
  - Triển khai project trong hệ thống Mater – Slave computers
- PySpark Streaming
  - Giới thiệu
  - Lý do chọn PySpark Streaming
  - Đặc điểm
  - Streaming Context/ DStream
  - Streaming Transformation Operations
  - Streaming Checkpoint
- GraphX
  - Giới thiệu
  - Làm việc với GraphX
    - Tạo graph
    - Vertex & edge
    - Trực quan Graph
    - Lọc thông tin trên graph (filtering)
    - Connecting
    - Tìm mối quan hệ (motif finding)
    - Đếm tam giác trên graph (triangle count)
    - Hạng trang (page rank)