**NỘI DUNG KHÓA HỌC KIẾN TRÚC BIG DATA**

**I. Mô tả khóa học**

Khóa học Kiến trúc Big Data cung cấp cho các cán bộ một nền tảng kiến thức cơ bản về Big Data frameworks để thiết kế và vận hành một Big Data Platform giải quyết các bài toán thực tế trong doanh nghiệp. Với cách tiếp cận dựa trên bài toán (problem-based) và dựa trên dự án (project-based) thực tế khóa đào tạo sẽ giúp cán bộ có được các kỹ năng thực hành và giải quyết các vấn đề thực tế sử dụng các kỹ thuật và công cụ của Big Data. Khóa học này gồm các nội dung như sau:

* Tổng quan về dữ liệu lớn, phân tích dữ liệu lớn, và các ứng dụng của nó
* Cài đặt và sử dụng thành thạo Apache Hadoop và HDFS
* Mô hình lập trình MapReduce – một nền tảng xử lý song song cho dữ liệu lớn
* Các framework Apache Sqoop và Apache NiFi trong thu thập Big Data
* Xử lý dữ liệu lớn với Apache Spark – Một công cụ mạnh nhất trong xử lý dữ liệu lớn hiện nay
* Xử lý luồng dữ liệu lớn (Streaming Big Data) với Spark Streaming & Kafka
* Kỹ thuật thiết kế kho dữ liệu lớn (Data Warehouses) trong Business Intelligence
* Thiết kế kiến trúc Trích xuất - Chuyển đổi - Nạp số liệu (ETL - Extract, Transform, và Load) trong Big Data Platforms
* Trực quan hóa dữ liệu với BIRT Report

**II. Công cụ sử dụng**

* Apache Hadoop, Apache Spark, Apache NiFi, Apache Hive, Apache Kafka, Birt Report
* Eclipse

**III. Nội dung**

Khóa học được chia thành 14 buổi + 1 buổi Seminar dự án cuối khóa học. Mỗi buổi ứng với một project gồm: 1.5 giờ Lý thuyết + 1.5 giờ Thực hành.

| **Bài** | **Tên bài** | **Mô tả** | **Thời lượng** | **Thời gian** | **Giảng viên** |
| --- | --- | --- | --- | --- | --- |
| 1 | **Tổng quan về phân tích dữ liệu lớn**  (Introduction to Big Data Analytics) | * Giới thiệu tổng quan về dữ liệu lớn * Các thách thức và cơ hội của phân tích dữ liệu lớn * Các công cụ và công nghệ trong phân tích dữ liệu lớn trong hệ sinh thái Hadoop * Thực hành: thiết lập một cụm Hadoop trên máy ảo | 3 | Sáng Thứ 5  30/12/2021 | TS. Nguyễn Văn Quyết |
| 2 | **Hệ thống tệp tin phân tán Hadoop**  (Hadoop Distributed File Systems) | * HDFS là gì? Vai trò của HDFS trong hệ thống Big Data. * Kiến trúc của HDFS * Quản lý các tệp tin trên HDFS qua Web Interface * Quản lý các tệp dữ liệu lớn trên HDFS qua Shell Commands * Thực hành tương tác với dữ liệu trên HDFS | 3 | Chiều Thứ 5  30/12/2021 | TS. Nguyễn Văn Quyết |
| 3 | **Xử lý dữ liệu lớn với MapReduce**  (Big Data Processing with MapReduce) | * Mô hình lập trình MapReduce * Quản lý tiến trình xử lý dữ liệu của Hadoop MapReduce qua Web Interface * Thực hành: xử lý dữ liệu lớn trên HDFS với Hadoop MapReduce | 3 | Sáng Thứ 5  06/01/2022 | TS. Nguyễn Văn Quyết |
| 4 | **Lưu trữ và Truy vấn dữ liệu lớn với Hive**  (Storing and Querying Big Data with Hive) | * Giới thiệu tổng quan về Apache Hive * Thao với với dữ liệu lớn có cấu trúc sử dụng Hive SQL * Thực hành: Thao tác với với dữ liệu lớn sử dụng Hive SQL | 3 | Chiều Thứ 5  06/01/2022 | TS. Nguyễn Văn Quyết |
| 5 | **Thu thập dữ liệu lớn (1)**  (Big Data Ingestion) | * Giới thiệu về Data Ingestion trong hệ thống Big Data * Data Ingestion từ nhiều nguồn sử dụng Apache NiFi * Thực hành: cài đặt Apache NiFi | 3 | Sáng Thứ 7  08/01/2021 | Chuyên gia Nguyễn Chí Thanh |
| 6 | **Thu thập dữ liệu lớn (2)**  (Big Data Ingestion Tools) | * Sử dụng các thành phần của NiFi:   + Flow file   + Processor   + Connector   + NiFi Repository * Thực hành: xây dựng luồng tích hợp dữ liệu sử dụng công cụ Apache NiFi   + Giới thiệu về xây dựng luồng NiFi   + Xây dựng luồng sử dụng RDBMS Connector   + Xây dựng luồng sử dụng Hdfs Connector | 3 | Chiều Thứ 7  08/01/2021 | Chuyên gia Nguyễn Chí Thanh |
| 7 | **Xử lý dữ liệu lớn với Spark (1)**  (Big Data Processing with Spark – 1) | * Tổng quan về Spark * Kiến trúc của Spark * Làm việc với Spark Core * Thực hành: sử dụng một số API quan trọng trong Spark Core trên Cloud | 3 | Sáng Thứ 5  13/01/2022 | Chuyên gia Nguyễn Chí Thanh |
| 8 | **Xử lý dữ liệu lớn với Spark (2)**  (Big Data Processing with Spark – 2) | * Giới thiệu Spark SQL * Truy vấn dữ liệu với Hive và Spark Thrift server. * Các định dạng dữ liệu trong Big Data Ecosystem * Giới thiệu về Data Frame * Sử dụng Data Frame Reader và Writer để đọc ghi dữ liệu * Thực hành: đọc ghi dữ liệu lớn trên Cloud với Spark | 3 | Chiều Thứ 5  13/01/2022 | Chuyên gia Nguyễn Chí Thanh |
| 9 | **Phân tích luồng dữ liệu lớn với Kafka & Spark (1)**  (Streaming Big Data Analytics using Kafka & Spark - 1) | * Giới thiệu về Kafka và vai trò trong hệ thống Big Data * Kiến trúc Kafka, các thành phần và cách hoạt động * Thực hành: cài đặt và sử dụng một cụm Kafka | 3 | Sáng Thứ 7  15/01/2022 | TS. Nguyễn Văn Quyết |
| 10 | **Phân tích luồng dữ liệu lớn với Kafka & Spark (2)**  (Streaming Big Data Analytics using Kafka & Spark - 2) | * Các mô hình xử lý dữ liệu dạng stream * Làm việc với Spark Streaming API * Thực hành: xây dựng một ứng dụng xử lý dữ liệu dạng stream đơn giản với Kafka và Spark | 3 | Sáng Thứ 7  15/01/2022 | TS. Nguyễn Văn Quyết |
| 11 | **Thiết kế kho dữ liệu trong BI**  (Designing a Data Warehouse in Business Intelligence) | * Các loại bảng trong hệ thống DWH * Kỹ thuật thiết kế bảng tổng hợp dữ liệu * Thực hành thiết kế bảng tổng hợp dữ liệu | 3 | Sáng Thứ 5  20/01/2022 | Chuyên gia Trần Đăng Hòa |
| 12 | **Thiết kế kiến trúc Trích xuất – Chuyển đổi – Nạp dữ liệu hiện đại**  (Designing  a Modular ETL (Extract, Transform, Load) | * Xây dựng luồng ETL với Pentaho * Phân biệt Relation, Job, Transform * Thực hành xây dựng luồng ETL xử lý file | 3 | Chiều Thứ 5  20/01/2022 | Chuyên gia Trần Đăng Hòa |
| 13 | **Thiết kế hệ thống báo cáo quản trị thông minh**  (Business Intelligence and Reporting Tools) | * Tổng quan về BI Applications. * Đưa ra vấn đề về nhu cầu số liệu thông qua các loại báo cáo khác nhau. * Giới thiệu công cụ BIRT cho Regular report kết hợp với hệ sinh thái Hadoop * Thực hành xây dựng báo cáo | 3 | Sáng Thứ 7  22/01/2022 | Chuyên gia Trần Đăng Hòa |
| 14 | **Thiết một một Big Data Platform**  (Design of a Big Data Platform) | * Thiết kế kiến trúc nền tảng xử lý dữ liệu lớn * Định cỡ tài nguyên và ước lượng đầu tư * Thực hành xây dựng một Big Data Platform | 3 | Chiều Thứ 7  22/01/2022 | Chuyên gia Trần Đăng Hòa |
| 15 | **Seminar dự án cuối khóa học** | | 3 | Sáng Thứ 5  10/02/2022 | TS. Nguyễn Văn Quyết |