deerwalk training center

DIPLOMA IN BIG DATA ANALYTICS WITH MONGODB

Nature of the Course: Theory + Practical

Total Hours per Day: 2 Hours

Course Duration: 2.5 Months + 1.5 Months (Internship)

Course Summary

The Diploma in Big Data Analytics is a comprehensive program focusing on the analysis and utilization of large and complex datasets. Students learn data collection, cleaning, integration, storage, and visualization techniques. They gain proficiency in programming languages, big data technologies, and advanced statistical and machine learning techniques. . It equips them with the skills and knowledge to extract valuable insights from big data and utilize them to drive business strategies and decision-making processes.

Completion Criteria

After fulfilling all of the following criteria, the student will be deemed to have finished the Module:

- Has attended 90% of all classes held.
- Has received an average grade of 80% on all assignments
- Has received an average of 60% in assessments.
- The tutor believes the student has grasped all of the concepts and is ready to go on to the next module.

Required Textbooks

- "Big Data Analytics: Methods and Applications" by S. Srinivasan.
- "Big Data Analytics with R and Hadoop" by Vignesh Prajapati.
- "Big Data Analytics with Spark and Hadoop" by Venkat Ankam.

Prerequisites

- Fundamental understanding of programming, bits/bytes, procedures, classes, and computer architecture. It's absolutely acceptable if you only have a theoretical understanding of programming, but you should be certain about what programming is and what you intend to gain from this session.
- If you are only interested in theory and have no interest/patience in spending at least 10 hours every week throughout the duration of the course, then this course might not be for you.
- If you have absolutely no idea about programming or do not see yourself doing programming in the

next six -odd months, then this class may not be for you.

Course Details

Basic Java Introduction And Concepts

First Month Spark

Week I

Overview Of Java Language

- Introduction
- Hardware and Software Requirements
- Installation of JDK

Programming With Java

- Class Declaration
- Members of Classes
- Structure of Java Classes
- Main Method
- Command Line Arguments
- Source Code Compilation
- Coding Convention

Constant, Variables And Data Types

- Primitive and Non-Primitive Variables

Week II

Decision And Branching

- If, Else, Switch, Break, Continue
- LOOPING
- For, While, Do-While

Fundamentals Of Loops

- Initializing Objects
- Static Members

- Inheritance
- Polymorphism
- Encapsulation

Week III

Abstract Class And Interfaces

- Defining Interfaces
- Separating Interface and Implementation
- Implementing and Extending Interfaces
- Abstract Classes

Exception Handling

- Exceptions and the Exception Hierarchy
- Throwing Exceptions
- Catching Exceptions
- Chaining Exceptions
- The Finally Block

Advance Data Structures (Java Collection Classes) Arrays

- List <e> Interface and its Implementation
- Map <k,v> Interface and Implementation
- Set <e> Interface and Implementation

Week IV

JDBC Connection

- JDBC Overview
- Using Driver Manager, Connection, Statement, Prepared Statement ● and Result Set
- Create, Delete, Insert, Update Statements

2nd Month Spark

Module 1

- Installing Spark
- Spark Architecture and RDDs
- Map and Reduces on RDDs
- Mapping and Outputting
- Tuples
- PairRdds
- FlatMaps and Filters
- Sorts and Coalesce
- Joins between 2 Rdds
- RDD Performance
- Exercise

Module 2

- Introduction to SparkSql
- Datasets
    - Datasets basics
    - Filters using expressions
    - Filters using columns
- Grouping and Aggregation
- Data Formatting
- Multiple Grouping
- Ordering
- DataFrame
    - Sql vs Dataframes
    - Dataframe Grouping
    - Dataframe joins
- More Aggregations in built functions
- User Defined functions
    - How to use a Lambda to write a UDF in Spark
    - Using more than one input parameter in Spark UDF
    - Using a UDF in Spark SQL
- SparkSQL Performance

    - Understand the SparkUI for SparkSQL
    - How does SQL and DataFrame performance compare?
    - Update - Setting spark.sql.shuffle.partitions
- HashAggregation
    - Explaining Execution Plans
    - How does HashAggregation work?
    - How can I force Spark to use HashAggregation?
    - SQL vs DataFrames Performance Results
- ETL Process for Big data Products Sales data and load the output in mongodb. ● Creating index in mongodb

3rd Month : Project Work

Front End

- Creating a Spring Boot Application with Login functionality
- Connection with mongodb’
- Integrate GRPC Api
- Creating a form to search product
- Display result
- Add field wise sorting and searching functionality
- Pagination

Backend End

- Learn about GRPC(Google Report Procedure Call)
- Use GRPC to build api
- Create all methods that needed for front end

Labs

Lab assignments will focus on the practice and mastery of contents covered in the lectures; and introduce critical and fundamental problem-solving techniques to the students.

Learning Outcomes

- Learn effective strategies for designing data models and schemas in MongoDB to optimize performance, scalability, and data integrity.
- Acquire skills in processing and analyzing big data using MongoDB's query language

- Develop a deep understanding of MongoDB's architecture, features, and functionalities for storing, querying, and manipulating large volumes of data.