Hive Technologies

Big Data Analysis

Course Outline (Course Duration: 2 Months)

1. Introduction to Big Data Analysis Course:

Begin your acquisition of Big Data knowledge with the most up-to-date definition of Big Data. You'll explore the impact of Big Data on everyday personal tasks and business transactions with Big Data use cases. Learn how Big Data uses parallel processing, scaling, and data parallelism. Learn about commonly used Big Data tools. Then, go beyond the hype and explore additional Big Data viewpoints.

2. Course Learning Objectives

CLO #	CLO Statement	Total Weeks
1	To familiarize with python basics	3
2	To understand the Data Analysis	1
3	To understand the Big Data Engineering	5

3. List of Software

Sr.no. #	Description	Payment Detail
1	Python setup	Open source
2	Java setup	Open source
3	Apache Spark	Open source
4	Hadoop	Open source
5	AWS	Free tier account

4. Lecture Breakdown:

Week	Lecture	Topic	Deliverables	CLO
1	1	Introduction to Python	• Introduction to python • Why python? • Google Colab for python coding • Python syntax and printing • Variables to perform various mathematical operations	1
	2	Python Data types (part 1)	• Built-in data types • Getting the data type and setting a specific data type • Strings operations i.e. slicing, concatenation, modification, formatting • List operations i.e. access list items, change list items, add or remove list, copy list, join lists	1
2	3	Python Data types (part 2)	• Tuples operations i.e. access tuple items, change tuples, unpack tuples, join tuples, tuple methods • Set operations i.e. access set items, add or remove set items, join sets, set methods • Python Dictionaries	1
	4	Conditional Statements and Loops	• Conditional Operators i.e. comparison operators, logical Operators, membership operators, identity operators • If-else, nested if • While loop, break or continue a while loop • For loop, break statement, continue statement, pass statement	1
3	5	Functions and Array	• Creating and calling functions, Lambda functions • Array operations i.e. access the elements of array, length of array, looping array, add or remove elements from array, array methods	1
	6	Classes/Objects, Math, JSON	• Creating and calling classes and objects • Inheritance and super functions • Built Math functions, Math module • Converting JSON to python or vice versa	1
4	7	File Handling	• Python File Handling i.e. read Files, write/Create Files, delete Files • Numpy Module	2
	8	Data Analysis	• Pandas introduction and file reading • Data cleaning and manipulation • Pandas plotting	2

5	9	Introduction to Big Data and PySpark (part 1)	• What is Big data? • Getting started with PySpark • Logistic Regression	3
	10	PySpark (part 2)	• Analysis using PySPark • Random Forests for Classification • Spark's MLlib	3
6	11	Database	• MySQL (create database, insert table, update, limit, join, order by) • NoSQL (create database, insert table, update, limit, join, order by)	3
	12	MongoDB	• Create database, insert, find, query, sort, join, limit, Data collection	3
7	13	Integration	• Integrate PySpark with Databases	3
	14	Pipeline and ETL (part 1)	• Basic ETL operations between databases and PySpark	3
8	15	Pipeline and ETL (part 2)	• Advanced ETL operations between databases and PySpark	3
	16	Pipeline and ETL (part 3)	• ETL operations between Cloud and PySpark	3
9	17	Hadoop and Hive SQL (Part 1)	• Apache Hadoop architecture, ecosystem, practices • Distributed File System (HDFS)	3
	18	Hadoop and Hive SQL (Part 2)	• MapReduce, HIVE and HBase. • Single node Hadoop cluster using Docker	3

5. Introduction to the Topics:

5.1. Introduction to Python:

This is chapter One where we explore what it means to write programs. Where, you will set things up so you can write Python programs. Moreover, in this first chapter we try to cover the "big picture" of programming so you get a "table of contents" of the rest of the book.

5.2. Python Data types (part 1):

In this lecture we will be covering Strings and Lists, and moving into data structures. As we want to solve more complex problems in Python, we need more powerful variables. Starting with lists we will store many values in a single variable using an indexing scheme to store, organize, and retrieve different values from within a single variable.

5.3. Python Data types (part 2):

The Python dictionary is one of its most powerful data structures. Instead of representing values in a linear list, dictionaries store data as key / value pairs. Using key / value pairs gives us a simple in-memory "database" in a single Python variable. Tuples are our third and final basic Python data structure. Tuples are a simple version of lists. We often use tuples in conjunction with dictionaries to accomplish multi-step tasks like sorting or looping through all of the data in a dictionary.

5.4. Conditional Statements and Loops:

Condition statement is a very simple concept - but it is how computer software makes "choices". Loops and iteration complete our four basic programming patterns. Loops are the way we tell Python to do something over and over. Loops are the way we build programs that stay with a problem until the problem is solved.

5.5. Functions and Array:

This is the lecture to learn about the implementations of the concepts we studied above with the help of functions and arrays.

5.6. Classes/Objects, Math, JSON:

In this lecture, we will create classes/object to learn the object oriented programming. We work with Application Program Interfaces / Web Services using the JavaScript Object Notation (JSON) data format.

5.7. File Handling:

In this lecture, we will train the students to handle Excel, CSV and JSON files. You can read and write these files to play with the datasets.

5.8. Data Analysis:

The lecture will involve all the elements of the specialization. In the first part of the capstone, students will do some visualizations to become familiar with the technologies in use and then will pursue their own project to visualize some other data that they have or can find.

5.9. Introduction to Big Data and PySpark (part 1):

Big Data as the digital trace that we are generating in this digital era. In this course, you will learn about the characteristics of Big Data and its application in Big Data Analytics. You will gain an understanding about the features, benefits, limitations, and applications of some of the Big Data processing tools. And basic into of the PySpark.

5.10. Introduction to Big Data and PySpark (part 2):

Apache Spark is an open-source processing engine that provides users new ways to store and make use of big data. It is an open-source processing engine built around speed, ease of use, and analytics. In this course, you will discover how to leverage Spark to deliver reliable insights. The course provides an overview of the platform, going into the different components that make up Apache Spark.

5.11. Database:

This lecture will introduce students to the basics of the Structured Query Language (SQL) as well as basic database design for storing data as part of a multi-step data gathering, analysis, and processing effort.

5.12. MongoDB:

This lecture will cover the emerging database technology i.e. NoSQL/MongoDB design for storing data as part of a multi-step data gathering, analysis, and processing effort.

5.13. Integration:

This lecture will cover the integration of the Apache Spark with the databases of MogoDB and MySQL to perform ETL operations to make data pipline to manage streaming data like stocks data.

5.14. Pipeline and ETL (part 1):

After integrating the Databases, we need to perform operations on databases using apache spark. This lecture will include the basic operations and concepts of ETL.

5.15. Pipeline and ETL (part 2):

This lecture will include the advanced operations of the databases with apache spark and perform high level of ETL methods to deal with streaming data .

5.16. Pipeline and ETL (part 3):

We use Cloud to make the data processing faster and secure. This lecture will be covering integration of the apache spark with cloud.

5.17. Hadoop and Hive SQL (Part 1):

Hadoop is an open-source framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. Hive, a data warehouse software, provides an SQL-like interface to efficiently query and manipulate large data sets residing in various databases and file systems that integrate with Hadoop.

5.18. Hadoop and Hive SQL (Part 2):

Hadoop concepts of MapReduce and HIVE will be covered in this lectures. This lecture will train the individuals to deal with implementation of big data concepts and streaming data.