Data Engineer – Training program-50 Hours

Python – 15 hours

Python Introduction

Python with Analytics, Bigdata, cloud

Tools & Utilities

Python Features

Python Environment setup – Installation, path

Python Execution – Interactive, Scripting, IDE

Keywords, Identifiers, Indentation

Variables & Datatypes

Operators – Arithmetic, Comparison, Assignment, Bitwise, Logical, Identity

Conditional statements – If, If Elif Else, Nested If

Loops – While, For & Nested

Control statements – Break, Continue, & Pass

List, Tuple & Dictionary

Functions – pass by value & reference, arguments, scope of variables

Modules – Importing, packages

Files I/O – Open, Close, Read, Write

Errors & Exceptions

OOPS - Classes & Objects

Regular Expression

XML/JSON processing

Lambda, Filter, Map, Reduce

Iterators, Decorators, Generators

Web scraping

Connection to Database, Fetch data

PIP

Unit Testing - pytest

Sys, OS, regex, csv, pyodbc, requests, json, NumPy, pandas, smtplib, pep8, pylint

Pyspark – 20 hours

Hadoop Vs Apace Spark

Spark Ecosystem & APIs

Spark Architecture

Spark Installation windows OS – Java JDK, Python, Spark with Hadoop

Spark session, Spark context, Spark conf, Spark shell

Resilient Distributed Dataset, Transformations & Actions, Data frames

Loading & Saving Data

Temporary table creation

Spark Web UI – Jobs, DAG

Coalesce & Repartition

SparkSQL - Aggregations – groupByKey, reduceByKey, combineByKey, aggregateByKey

Spark - streaming, GraphX, MLib

Spark submit command

Monitoring Execution, Submitting Job

Databricks setup

AWS – 5 hours

AWS EC2 Pyspark setup

AWS EMR cluster setup

S3, SNS

Airflow – 5 hours

Airflow framework, DAG, UI & Webserver

Operators, Plugins, Hooks, connections & variables

Visualizing, scheduling, Series & Parallel Jobs

Resume Preparation – 2 hours

Interview Guidance – 3 hours

All code snippets discussed in class will be shared in GitHub Repo

FAQ's document will be provided for interview preparation