Python, AWS and PySpark

Syllabus

1: Introduction to Python

- Installation and Working with Python
- Introduction, why python?
- Versions of Python
- SET PATH
- PEP 8 standards
- Coding conventions
- Understanding Python variables
- Identifier rules
- Literals
- Keywords
- IDLE and information
- Different ways of execution
- Scripting
- Python Operators
- Understanding python blocks
- Indentation, comments, docstring
- Type casting, Unicode etc.

2: Python Data Types

- Mutable and Immutable data types
- Declaring and using Numeric data types: int, float, complex.
- Using string data type and string operations

- Defining list and list slicing, its methods
- Use of Tuple data type

3: Python Program Flow Control

- Conditional blocks using if, else and elif
- Nested if, elif ladder
- Simple for loops in python
- For loop using range, string, list and dictionaries.
- Use of while loops in python
- Loop manipulation using: pass, continue, break
- Programming using Python conditional and loops block.
- Different case studies

4: Python String, List, set and Dictionary Manipulations

- Building blocks of python programs
- Understanding string built-in methods
- List manipulation using built-in methods
- Tuple operation
- Set: its methods and manipulation
- Dictionary: its methods and manipulation
- Functions
- Modules and Packages

5: Fundamentals of Object orientation:

- What is OOP
- Class
- Reference variable
- Types of variables
- Types of Methods
- Importing Class
- Constructor
- OOP's Concepts: Inheritance, Encapsulation, Polymorphism, Abstraction
- File handling in detail: txt, bin, csv

6: Data Engineering

- Core Concepts
- ETL
- Data Architectures on AWS
- AWS Data Engineering Services

7: AWS Cloud

- What is Cloud Computing?
- The Deployment Models of the Cloud,
- Types of Cloud Computing.
- AWS Regions.
- AWS Availability Zones
- Shared Responsibility Model diagram
- AWS load balancing and Auto Scaling

8: AWS Snow Family

- Data migration
- Edge computing
- Snow Family – Edge Computing

9 : Cloud Integration Section

- Amazon SQS
- Amazon SNS

10 : AWS Machine Learning

- Amazon Rekognition
- Amazon Transcribe
- Amazon Lex & Connect
- Amazon Comprehend
- Amazon SageMaker
- Amazon Forecast
- Amazon Textract

11 : Account Management, Billing & Support Section

- Multi Account Strategies
- AWS Organization
- Service Control Policies (SCP)
- SCP Hierarchy
- Blacklist and Whitelist strategies
- Pricing Models in AWS
- Free services & free tier in AWS
- Compute Pricing
- Storage Pricing
- Database Pricing
- Networking Costs
- Savings Plan, Cost Explorer
- AWS Support Plans Pricing

12 : AWS Step Functions

- Build serverless visual workflow to
- orchestrate to Lambda functions.

13: AWS S3

- Create S3 Bucket and Upload Content to Bucket and Manage their Access
- Enable S3 Encryption S3 Versioning
- Lifecycle management rules on S3 Bucket
- Objects and Buckets in AWS S3
- Amazon S3 – Security
- Amazon S3 - Versioning

14 : AWS Glue

- AWS Glue Data Catalog
- AWS Glue Databases
- AWS Glue Tables
- AWS Paritions
- AWS Glue Crawler
- Create Table In The Glue Data Catalog
- AWS Glue Connections
- AWS Glue ETL Job
- AWS Glue Triggers
- AWS Glue Dev Endpoints

15 : AWS Lambda

- Lamda Console
- Lambda Demo
- Create Function
- Bluprint Configuration
- Viewing Logs
- IAM Permissions

16 : Parquet file

- Introduction
- Jupyter Notebook
- Pyarrow installation
- Data execution/ Reading Parquet file using Python-pandas
- Convert parquet to csv in Python with Pandas
- Write parquet with Partitions to AWS S3

17 : PySpark with Python

- PySpark Dataframe
- Reading The Dataset
- Checking the Datatypes of the Column (Schema)
- Selecting Columns And Indexing
- Check Describe option similar to Pandas
- Adding Columns
- Dropping columns

- Renaming Columns
- Pyspark Handling Missing Values
- Pyspark GroupBy And Aggregate Functions
- Cover around 30 Spark Functions

18 : PySpark for AWS Glue

- Spark And PySpark Theory
- Spark DynamicFrame
- Spark DynamicFrame Manipulation and Transformations
- Spark DataFrame Theory
- Spark DataFrame Manipulation and Transformations

19 : AWS Glue Studio

- Setting Up The Sample Data
- Configuring The IAM Role
- Setting Up The Crawler

20 : AWS Crawlers and Athena

- Create a database pipeline
- Load Database
- Data Analysis

21 : AWS Kinesis

- Set Up Work
- Kinesis Streams Theory
- SDK Vs KPL Theory
- Kinesis Data Streams Practical
- Kinesis SDK
- KPL Practical
- Lambda Consumer Practical
- KCL Practical
- Studio Note Book Practical
- Kinesis Firehose Practical
- Kinesis Analytics Practical

22 : AWS add-ons

- AWS EMR Serverless
- AWS Security using IAM overview
- AWS CLI overview
- AWS Redshift

  - Redshift Tables

  - Redshift Architecture

23 : pySpark Functions and Transformations

- Read / Write Dataframe using pyspark
- Read / Write Dataframe into json file using pyspark
- Read / Write Dataframe into parquet file using pyspark
- join() function in PySpark | inner, left, right, full Joins
- join() Left semi, Left anti & self join
- Convert RDD to Dataframe in PySpark
- structType() / structField()
- explode(), split(), array() & array_contains()
- MapType Column
- ArrayType columns
- Column Class
- Row() class in PySpark
- when() & otherwise() functions

24 : AWS EC2

- Introduction
- What is AWS EC2?
- Why use AWS EC2?
- When use AWS EC2?
- EC2 tutorial on the AWS Console

25 : AWS Cloudformation

- Introduction
- What is cloudformation?
- Why use cloudformation?
- When do we use cloudformation?
- Hands on tutorial

26 : AWS Iceburg

- Iceberg Theory
- Set Up Work
- Apache Iceberg Tutorial
- Create CSV table
- Create Iceberg table
- Insert Data Into Iceberg table
- Select Data From Iceberg table
- Updata Data
- Time Travel In Iceberg
- Delete Data
- Clean Up Resources

26 : Project –

Create relational and NoSQL data models using real-time dataset and perform PySpark transformation using Data Engineering concept.