PYTHON WITH DATA ENGINEERING STACK

Course Curriculum

TRAINERS

A. JOHN PAUL ANTONY &
G.S.RAMAN

LEARNING & DEVELOPMENT TEAM

Python with Data Engineering

Tech Stack	Total Number of Trainees
PYTHON WITH DATA ENGINEERING	6
TRAINER NAME	**2023 Batch I**
A. JOHN PAUL ANTONY & G.S. Raman	

S.No	Name of the Topics	Topics
	Introduction	
1	Introduction	**1. Programming Paradigms** What is paradigm? Common Programming Paradigms Imperative Programming Structured Programming Object Oriented Programming Declarative Programming Functional Programming Aspect oriented programming Event-driven programming Real world Examples
		2. Data Structures and Algorithms What is a data structure? Classifying data structures Commonly used data structures Collections Operations on data structures Search Sort Iterate/fetch Add Remove Space and Time complexity
		3. Linux & Networking Basics of Operating System, Types of OS , Linux Basic and advanced Commands ,Linux distributions, Networking Commands, Computer Network Basics - IP address, Port, Ethernet, MAC and OSI layer. Computer Network Protocols, Wi-Fi, HTTP, unicast, broadcast, multicast, GSM, Server Client architecture, Topology, Network Devices, Firewall, Virus, types of Virus,

		Antivirus, Basics of Cryptography, Domain and Workgroup, Social Networks.
MySQL & MongoDB Databases		
2	**Databases**	**1. MySQL -** RDBMS Concepts , ER Diagram DDL , DML & DCL , TCL RDBMS Normalization ACID Support Cardinality Popular RDBMS SQL Basics, SQL Data Types, Operators, SQL Commands- DDL, DML, DCL, TCL, Constraints, Keys, Normalization, SQL Joins, Views, Stored Procedure, Trigger, Cursor, SQL functions Database Schema Design – Schema Modelling Real world examples
		2. MongoDB - Introduction to NoSQL, Create, Insert, Sharding in Mongo DB , Working in Atlas(cloud), Update, Save, Remove
Python Core & Advanced Python		
3	**Core Python**	1. Introduction & Python Setup and Data types
		2. Python Objects and Data structure Basics & Advanced
		3. I/O Formatting, Operators
		4. Control Structures - Statements, Looping
		5. Patterns & Data Manipulations - String, List, Tuple, Sets, frozen sets and Dictionary
		6. Advanced List processing, List Comprehensions, Aliasing, Cloning, Pickling and unpickling and Dictionary Comprehensions
		7. Functions - Built in functions, Module functions, user defined functions and Concepts
		8. Useful functions - range, Zip, enumerate, all, any, Lambda, Map, reduce, filter
		9. *args, **kwargs & Methods

		10. Modules and Packages , __name__ and __main__ & Modules Attributes
4	**Advanced Python**	1. Object Oriented Programming - Introduction, Class, Object, Class attributes, __init__,self, Methods, Inheritance and polymorphism, Magic Methods, Abstract Class, Encapsulation, Overloading, Super, Adding and retrieving Dynamic attributes, Duck Type, Method Resolution Order(MRO), doc_string, Overloading, overriding
		2. File Handling - Modes of Operations, File Methods, Working with CSV & PDF Files
		3. Errors and Exception Handling
		4. Iterators, Itertools, Infinite, terminatoric, combinatoric Iterator functions
		5. Generators and Decorators
		6. Thread and Multithreading
		7. Network Programming & Simple Chat application Programming
		8. Python Database Connectivity - MySQL DB, CRUD Operations
		9. Design Pattern – Creational, Structural, Behavioral and Architectural Pattern – Singleton, Decorator, factory, Iterator, Mediator and MVC pattern
		10. Console Application with MVC Pattern and Coding Standards
	Testing	Unit test framework and PyTest framework – Test case, Test Suite, Test fixture, Test loader, Test Runner Class. Assert Functions etc.,
Python Web Application Development		
6	**Web Design**	1. Browser Working & Configuration Model
		2. HTML5 - HTML Basic Elements and attributes & values, HTML Vs HTML5, HTML5 Form elements and attributes
		3. CSS3 - Types of Styles, Selectors, Basic Properties, Box Model, Flexbox Layout, Display , Table Properties, Media Queries, Animations

		4. JS6 - Functions, DOM Structure, DOM and BOM Methods, Objects, Form Validations
		5. Web Server , Database Server and File Server and Domain Name Creation with free of cost, Hosing Websites into Server
7	**Source Code Management tool**	1. Introduction to Devops, Git and GitHub Configuration, Centralized and distributed VCS Concepts, Git Commands , Realtime SCM Scenario
8	**Web Framework–Django 4**	1. Web Server and Web Application Working Principle
		2. Python Web Frameworks - Django Vs Flask
		3. Django Framework Setup Configuration, Virtual Environment Configuration and project Structure and Django Server Configuration
		4. Django Architectural overview , MVT Pattern
		5. URL Mapping or routing, Views, Templates, Template inheritance
		6. URL Path and Rendering in views , DTL and Jinja2 Template Engine
		7. DTL - Jinja2 Syntax and Passing dynamic data into html template
		8. Models Creation, Databases, Migrations, ORM and Model object methods, Lookups, Re-migration
		9. ORM Relational Mapping, Population Scripts, Relative URL with templates
		10. Template filters and custom filters , Static files
		11. Django Forms, Form Classes, Form Validation, Model Forms, CRUD Django functions in Form.
		12. Django Authentication & Authorization, User Models, User Models and Forms Applications MVT pattern
		13. Registration, Login & Logout function
		14. Advanced Topics – CBV’s, Template views with CBV, Detail View and List

		View, CRUD Views 15. Customizing the Django Admin – Admin Panel, Superuser creation, admin templates, ordering fields, adding search, adding filters, adding fields, editable List View 16. Django Debug Toolbar, Bootstrap 4 – Buttons, Forms, Navbar, Grids, Layout, Card, Carousel, Layouts 17. Django Deployment – GitHub and Git Configuration and Full deployment walkthrough on python anywhere. 18. Sample Web application projects and deployment into Python anywhere, AWS, Heroku platform.
9	**Rest API - Django**	1. API View, Configure View URL, Create a Serializer, POST Function, add PUT, PATCH, DELETE Methods and test the methods. 2. Viewset, URL Router, add Create, retrieve, update, partial_update and destroy functions, Test Viewset. 3. Profile API, login API, Profile feed API, Deploying API to AWS and Pythonanywhere cloud platform.
AWS Infrastructure Automation – Python Script		
9	**AWS Boto3 Script**	1. AWS – Python Boto3 Script for EC2, IAM, S3 and CloudWatch, SNS Services. Introduction to Lambda function in AWS.
Data Analysis & Processing Tools		
10.	**Python Library - Data Analysis & Visualization Tools**	1. Numpy – Scientific Computation 2. Pandas - Tabular data 3. Scikit Learn – Data Modelling & Preprocessing 4. Matplotlib – Data Visualization 5. Case Study & Real time Use cases Implementation
11.	**PySpark - Introduction**	Setup Development Environment (Windows 10) - Introduction Setup Development Environment - Python and Spark - Pre-requisites Setup Development Environment - Python Setup on Windows Setup Development Environment - Configure Environment Variables Setup Development Environment - Setup PyCharm for developing Python applications

		Setup Development Environment - Pass run time arguments or parameters Setup Development Environment - Pass run time arguments or parameter Setup Development Environment – Install 7zip for uncompress and untar on windows Setup Development Environment - Setup Spark Setup Development Environment - Install JDK Setup Development Environment – Configure environment variables for Spark Setup Development Environment – Install WinUtils – integrate Windows and HDFS Setup Development Environment – Integrate PyCharm and Spark on Windows 10
12	**PySpark - Apache Spark - Transform, Stage and Store**	Introduction Introduction to Spark Setup Spark on Windows Quick overview about Spark documentation Connecting to the environment Initializing Spark job using pyspark Create RDD from HDFS files Create RDD from collection - using parallelize Read data from different file formats - using sqlContext Row level transformations - String Manipulation Row Level Transformations - map Row Level Transformations - flatMap Filtering data using filter Joining Data Sets - Introduction Joining Data Sets - Inner Join Joining Data Sets - Outer Join Aggregations - Introduction Aggregations - count and reduce - Get revenue for order id Aggregations - reduce - Get order item with minimum subtotal for order id Aggregations - countByKey - Get order count by status Aggregations - understanding combiner Aggregations - groupByKey - Get revenue for each order id groupByKey - Get order items sorted by order_item_subtotal for each order id Aggregations - reduceByKey - Get revenue for each order id Aggregations - aggregateByKey - Get revenue and count of items for each order id

		Sorting - sortByKey - Sort data by product price Sorting - sortByKey - Sort data by category id and then by price descending Ranking - Introduction Ranking - Global Ranking using sortByKey and take Ranking - Global using takeOrdered or top Ranking - By Key - Get top N products by price per category - Introduction Ranking - By Key - Get top N products by price per category python collections Ranking - By Key - Get top N products by price per category - using flatmap Ranking - By Key - Get top N priced products - Introduction Ranking - By Key - Get top N priced products - using Python collections Ranking - By Key - Get top N priced products - Create Function Ranking - By Key - Get top N priced products - integrate with flatMap Set Operations - Introduction Set Operations - Prepare data Set Operations - union and distinct Set Operations - intersect and minus Saving data into HDFS - text file format Saving data into HDFS - text file format with compression Saving data into HDFS using Data Frames - json
13	**PySpark - Apache Spark 1.6 - Data Analysis - Spark SQL or HiveQL using Spark Context**	Different interfaces to run SQL - Hive, Spark SQL Create database and tables of text file format - orders and order_items Create database and tables of ORC file format - orders and order_items Running SQL/Hive Commands using pyspark Functions - Getting Started Functions - String Manipulation Functions - Date Manipulation Functions - Aggregate Functions in brief Functions - case and nvl Row level transformations Joining data between multiple tables Group by and aggregations Sorting the data Set operations - union and union all Analytics functions - aggregations Analytics functions - ranking Windowing functions Creating Data Frames and register as temp tables

		Write Spark Application - Processing Data using Spark SQL Write Spark Application - Saving Data Frame to Hive tables Data Frame Operations
14	**PySpark - Setup Hadoop and Spark Environment for Practice**	Introduction to Setting up Environment for Practice Creating Virtual Machine Starting HDFS and YARN Gracefully Stopping Virtual Machine Understanding Datasets provided in Virtual Machine
15	**PySpark - Apache Spark 2.x - Data processing - Getting Started**	Introduction Review of Setup Steps for Spark Environment Quick Review of Spark APIs Spark Modules Spark Data Structures - RDDs and Data Frames Develop Simple Application Apache Spark - Framework
16	**PySpark - Apache Spark 2.x - Data Frames and Pre-Defined Functions**	Introduction Data Frames - Overview Create Data Frames from Text Files Create Data Frames from Hive Tables Create Data Frames using JDBC Data Frame Operations - Overview Spark SQL - Overview Overview of Functions to manipulate data in Data Frame fields or columns
17	**PySpark - Apache Spark 2.x - Processing Data using Data Frames - Basic Transformations**	Define Problem Statement - Get Daily Product Revenue Selection or Projection of Data in Data Frames Filtering Data from Data Frames Joining multiple Data Frames Perform Aggregations using Data Frames Sorting Data in Data Frames Development Life Cycle using Data Frames Run applications using Spark Submit Data Frame Operations - Window Functions - Overview Data Frames - Window Functions APIs - Overview Define Problem Statement - Get Top N Daily Products Data Frame Operations - Creating Window Spec Data Frame Operations - Performing Aggregations using sum, avg etc Data Frame Operations - Time Series Functions such as Lead, Lag etc

		Data Frame Operations - Ranking Functions - rank, dense_rank, row_number etc
18	**PySpark - Apache Spark using SQL - Getting Started**	Getting Started - Overview Launching and using Spark SQL CLI Overview of Spark SQL Properties Running OS Commands using Spark SQL Understanding Warehouse Directory Managing Spark Metastore Databases Managing Spark Metastore Tables Retrieve Metadata of Tables Role of Spark Metastore or Hive Metastore Exercise - Getting Started with Spark SQL
19	**PySpark - Apache Spark using SQL - Basic Transformations using Spark SQL**	Introduction Hands on Project, Filter, Join, Aggregate and sorting Data & DDL and DML Creating Tables using Parquet Load vs. Insert Inserting Data using Stage Table Creating Partitioned Tables Adding Partitions to Tables Loading Data into Partitioned Tables Inserting Data into Partitions Using Dynamic Partition Mode Exercise - Partitioned Tables
20	**PySpark - Apache Spark using SQL - Pre-defined Functions**	Introduction - Overview of Spark SQL Functions Overview of Functions Validating Functions String Manipulation Functions Date Manipulation Functions Overview of Numeric Functions Data Type Conversion Dealing with Nulls Using CASE and WHEN Query Example - Word Count
21	**PySpark - Apache Spark SQL - Windowing Functions**	Introduction to Windowing Functions Prepare HR Database Overview of Windowing Functions Aggregations using Windowing Functions Using LEAD or LAG

		Getting first and last values Ranking using Windowing Functions Order of execution of SQL Overview of Subqueries Filtering Window Function Results
22	**PySpark – Case study**	Guidelines for initializing the job Assessment
AWS Essential & Analytical Services		
23	**Introduction**	Introduction to basic terminologies of cloud and AWS Pre requisites
24	**AWS Essential Services**	EC2, S3, IAM, SNS, SQS, Cloud Watch, VPN
25	**Amazon Kinesis Data Streams and Firehose**	AWS Kinesis Streams and Kinesis Firehose Basics AWS Kinesis Firehose - Create a Delivery Stream AWS Kinesis Firehose - Post data to a delivery stream AWS Kinesis Firehose - Delete Delivery Stream AWS Kinesis - Create a Kinesis Data Stream AWS Kinesis - Create a Kinesis Data Stream Consumer AWS Kinesis - Create a Kinesis Data Stream Producer AWS Kinesis - Post data to a Kinesis Delivery Stream AWS Kinesis Data Streams - Delete Streams
26	**Amazon Managed Streaming for Kafka**	AWS MSK - Architecture Diagram, Use-Case AWS MSK - Create a network for hosting brokers AWS MSK - Create Kafka Cluster using MSK AWS MSK - Create a Kafka Client to connect to MSK Kafka Cluster AWS MSK - Delete Kafka Cluster Instance
27	**AWS Database Migration Service (DMS) and Schema Conversion Tool(SCT)**	AWS DMS and SCT - Architecture Diagram, Use-Case AWS SCT - Analyze Relational Database Schema AWS SCT - Create a Redshift Cluster as destination AWS SCT - Compare schema mapping from SQL Server OLAP to RedShift AWS SCT - Assess schema objects for conversion AWS SCT - Apply Schema migration to Redshift Data Warehouse AWS SCT - Delete SQL Server and Redshift Instance AWS DMS - Create a DMS Target AWS DMS - Create a DMS Source AWS DMS - Configure DMS Source AWS DMS - Configure Security and Data Structures AWS DMS - Replication Instance and Database endpoints

		AWS DMS - Create Migration or Replication Task AWS DMS - Delete DMS Instance and other resources
28	**AWS Data Sync and Storage Gateway**	AWS Data Sync - Architecture Diagram, Use-Case AWS Data Sync - Create an Agent AWS Data Sync - Configure an Agent AWS Data Sync - Create a data transfer Task AWS Data Sync - Execute a data transfer task AWS Data Sync - Delete agent and tasks AWS Data Sync - Comparison with Storage Gateway
29	**Amazon S3 and Glacier - Advanced**	Amazon S3 - Transfer Acceleration Amazon S3 - Storage Tiers and Life-cycle Management Rules Amazon S3 - Cross Region Replication Amazon S3 - Storage Analytics, Usage Metrics, and Inventory reports Amazon S3 - Object Locking Amazon S3 - S3 Select for Big Data Amazon Glacier - Create an Archive Vault Amazon Glacier - Move data from S3 to Glacier Amazon Glacier - Retrieve data from Glacier Amazon Glacier - Capacity Units and Glacier Select
30	**Amazon (AWS) RDS - MySQL**	AWS RDS MySQL - Basic and Advanced Settings AWS RDS MySQL - Querying database instance using MySQL Workbench AWS RDS MySQL - Performance Insights AWS RDS MySQL - Create Read Replicas AWS RDS MySQL - Test Read Replica AWS RDS MySQL - Create Aurora Read Replica AWS RDS MySQL - Delete Master and Replica Instance
31	**Amazon (AWS) RDS - PostgreSQL**	AWS RDS PostgreSQL - Basic and Advanced Settings AWS RDS PostgreSQL - Querying database instance AWS RDS PostgreSQL - Create Read Replica AWS RDS PostgreSQL - Promote Read Replica AWS RDS PostgreSQL - Delete Instance
32	**Amazon DynamoDB, API Gateway and Lambda**	AWS DynamoDB - Architecture and Use-Cases AWS DynamoDB - Integration AWS DynamoDB - Download DynamoDB Local Edition AWS DynamoDB - Create a Table with Partition Keys and Add Data AWS DynamoDB - Fundamentals, Partitioning and Indexing AWS DynamoDB - Global Secondary Index, Local Secondary Index

		AWS DynamoDB - Table Properties and Features Walkthrough AWS DynamoDB - Backups, Reserved Capacity and Preferences AWS DynamoDB - Create API Gateway and Lambda Function AWS DynamoDB - Configure Lambda Function AWS DynamoDB - Configure API Gateway and IAM Role AWS DynamoDB - Code lambda function to read and write data AWS DynamoDB - DynamoDB Streams Architecture and Use-Cases AWS DynamoDB - Create DynamoDB Stream, Trigger and Lambda function AWS DynamoDB - Code Lambda function and Test DynamoDB Streams AWS DynamoDB - Enable and Configure TTL Feature AWS DynamoDB - Global Database and On-Demand Capacity provisioning AWS DynamoDB - DAX (DynamoDB Accelerator)
33	**Amazon Data Pipeline**	AWS Data Pipeline - Architecture Diagram, Use-Case AWS Data Pipeline - Create source and destination data repositories AWS Data Pipeline - Create a new data pipeline AWS Data Pipeline - Execute data pipeline AWS Data Pipeline - Delete data pipeline
34	**Amazon Redshift and Redshift Spectrum**	Amazon Redshift - Introduction Amazon Redshift - Architecture Amazon Redshift - Clustering Sizing Amazon Redshift - Network Configuration Amazon Redshift - Create Redshift Cluster Amazon Redshift - Setup Redshift Client and access Redshift Cluster Delete Redshift Cluster Amazon Redshift Spectrum - Architecture
35	**Amazon Elasticsearch**	AWS ElasticSearch - Use-Cases, Basics of ElasticSearch AWS ElasticSearch - Architecture Diagram & Pricing AWS ElasticSearch - Create ElasticSearch domain AWS ElasticSearch - Explore properties AWS ElasticSearch - Query domain with ElasticSearch Client AWS ElasticSearch - Delete domain
36	**Amazon Elastic Map Reduce - AWS EMR**	AWS EMR - Cluster Architecture Diagram and Details AWS EMR - Storage Architecture Diagram and Details AWS EMR - Integration Architecture Diagram and Details AWS EMR - Setup and Install Hive JDBC Driver AWS EMR - Create cluster using Quick Create Options AWS EMR - Clone an EMR cluster

		AWS EMR - Query DynamoDB from EMR using Hive and SQL Workbench AWS EMR - Terminate EMR Cluster AWS EMR - Setup Glue Catalog integration with Hive AWS EMR - Create data structures in Glue Catalog from Hive AWS EMR - Setup R Studio AWS EMR - Access data from S3 in R Studio AWS EMR - Setup VPN Tunnel to EMR AWS EMR - Overview of accessing EMR Framework User Interface AWS EMR - Access S3 Data with Hue, Hive, Tez AWS EMR - Oozie Overview AWS EMR - Create a Notebook Cluster AWS EMR - Access S3 Data in EMR Notebook with PySpark, SparkR and Python AWS EMR - Terminate Notebook Cluster
37	**Amazon Backup**	AWS Backup - Architecture, Use-Cases AWS Backup - Create Backup Plan and Backup Rule AWS Backup - Assign resources to backup plan AWS Backup - Create On-Demand and Scheduled Backup Jobs AWS Backup - Delete Resources
38	**AWS Glue**	AWS Glue - Introduction AWS Glue - Architecture AWS Glue - Add Table manually to Data Catalog AWS Glue - Create and execute a Crawler AWS Glue - Crawl different datastores in a single crawler job AWS Glue - Create and execute a ETL Job AWS Glue - Convert Data to Parquet format AWS Glue - Create Trigger for a Glue Job
39	**AWS Athena**	AWS Athena - Introduction AWS Athena - Creating tables and querying data AWS Athena – Limitations
40	**AWS QuickSight**	AWS QuickSight - Introduction AWS QuickSight - Overview AWS Quicksight - Architecture AWS Quicksight - Setup and Pricing AWS Quicksight - Components and Reporting Authoring Workflow AWS Quicksight - Visualizations AWS Quicksight - Author a Quicksight Report

41	**Amazon Lake Formation**	Preview of Data Lake formation
42	**Case Study**	Data Analysis using AWS Services + Amazon SageMaker Notebook
Devops Tools		
43	**DevOps Tools**	Git, GitLab, Jenkins, Maven, Chef, Ansible, Docker and Kubernetes