**PYSPARK**

What is Big Data

Big Data Analytics

Huge Amount of Data

Heterogeneous Data

Accessing processing speed

5Vs of Big Data (Velocity, Volume, Variety, Veracity, Value)

What is Hadoop?

Apache Hadoop software is an open source framework that allows for the distributed storage and processing of large datasets across clusters of computers using simple programming models.

Hadoop is designed to scale up from a single computer to thousands of clustered computers, with each machine offering local computation and storage. In this way, Hadoop can efficiently store and process large datasets ranging in size from gigabytes to petabytes of data.

* Hadoop is a framework which is used to store Big Data across a spectrum of devices
* This is done to help you process Big Data in parallel
* The important components in the Hadoop Ecosystem that you should know: -

1. HDFS ---- Namenode and Datanode
2. MapReduce ----- Replaced by PySpark
3. YARN

**Introduction to Apache Spark**

Spark was introduced by **Apache Software Foundation** for speeding up the computational procedures of the software processes.

Spark uses Hadoop in two ways one is **storage**, and another is **cluster management** computation, it uses Hadoop for **storage purpose** only

**What is Apache Spark?**

* Apache Spark is a **lightning-fast cluster computing** technology designed for fast computation
* It is based on Hadoop **MapReduce** and it extends the **MapReduce** model **to efficiently** use it for more types of **computations**, which includes interactive queries and **stream processing**
* Spark is one of Hadoop’s sub-project developed in **2009** in UC Berkeley’s **AMPLab** by Matei Zaharia

**Features of Apache Spark**

* Speed
* Supports multiple languages
* Advanced Analytics
* In-Memory Computations
* Fault Tolerance
* Lazy Evaluation

**Apache Spark Ecosystem**

Apache Spark Core: -

* Spark Core is the underlying general execution engine for Spark platform that all other functionality is built upon
* It provides In-Memory computing and referencing datasets in external storage systems
* Spark SQL
* Spark GraphaX
* Spark Streaming
* MLlib

**PySpark** is an API written in Python to support **Apache Spark**

Apache Spark is **a distributed framework** that can handle Big Data Analysis