TECHGENX

Spark and Python for Big Data with PySpark course

Please Note: All below course content will be covered in practical scenarios and regular assignments will be shared. Along with the below course, around 100+ programs will be shared for self-practice

- What is Big Data?
- Hadoop and HDFS?
- Apache Spark?
- Setting up Python with PySpark.
    - Use Python and Spark together to analyze Big Data
    - Work on Consulting Projects that mimic real world situations!
- Databricks setup
    - Learn about the DataBricks Platform!
- Local VirtualBox set up
- AWS EC2 Pyspark setup
    - Get set up on Amazon Web Services EC2 for Big Data Analysis.
    - Learn how to use AWS Elastic MapReduce Service!
- AWS EMR Cluster set up
- Spark dataframe Basics.
    - What is Spark Session and how to use it
    - Creating dataframes using PySpark
    - Reading and writing Csv files
    - Reading Json files using PySpark
    - Data Fabrication using PySpark
    - Join 2 data frames in pyspark
    - Udf functions in pyspark (replacement of pandas lambda function)
    - Use of spark-submit command in terminal for PySpark Application
    - Customize logger lib of PySpark as per project in real world scenario.
    - Creating schemas and adopting user-defined schemas in PySpark.
    - Use of function lib in Pyspark.
- Spark Dataframe Project exercise
- Introduction with Machine learning with Mlib
    - Use Spark's MLlib to create Powerful Machine Learning Models
- Linear Regression
- Logistics Regression
    - Classify Customer Churn with Logisitic Regression
- Decision Tree and Random Forests
    - Learn how to use Spark's Gradient Boosted Trees

    - Use Spark with Random Forests for Classification
- K-means clustering
- Collaborative Filtering for Recommender Systems
- Natural Language processing
    - Create a Spam filter using Spark and Natural Language Processing!
- Spark streaming with Python
    - Learn how to leverage the power of Linux with a Spark Environment.
    - Use Spark Streaming to Analyze Tweets in Real Time!