Full Stack Databricks and Data Science using PySpark
By Dr. Vishwanath Rao

Topic 1: Introduction to Databricks and PySpark

1: Overview of Databricks and PySpark
- Introduction to Databricks platform and its features
- Overview of PySpark and its capabilities
- Setting up a Databricks environment and getting started with PySpark

2: PySpark Basics
- Basics of PySpark DataFrame and RDD APIs
- Data ingestion using PySpark
- Basic data exploration and manipulation with PySpark

3: Data Preprocessing with PySpark
- Handling missing values and outliers
- Data transformation and feature engineering with PySpark
- Exploratory Data Analysis (EDA) using PySpark

Topic 2: Advanced Data Processing with PySpark

4: Advanced PySpark Transformations
- Advanced data transformations and aggregations
- Window functions and partitioning in PySpark
- Handling complex data types with PySpark

5: PySpark MLlib for Machine Learning
- Introduction to PySpark MLlib for machine learning
- Building machine learning pipelines with PySpark
- Model training, evaluation, and tuning using PySpark MLlib

6: Databricks SQL and Data Visualization
- Introduction to Databricks SQL for querying and analyzing data
- Data visualization with Databricks notebooks
- Creating interactive dashboards using Databricks

Topic 3: Real-time Data Processing and Advanced Topics

7: Real-time Data Processing with Structured Streaming
- Overview of structured streaming in PySpark
- Building real-time data pipelines with Structured Streaming

- Handling streaming data and window operations with PySpark

8: Performance Tuning and Optimization
- Strategies for optimizing PySpark performance
- Understanding query optimization and execution plans
- Practical tips for improving PySpark job performance

9: Scalable Machine Learning with PySpark
- Handling large datasets and distributed computing with PySpark
- Advanced machine learning techniques with PySpark MLlib
- Case studies and examples of scalable machine learning projects

Topic 4: Project Work and Showcase

10-14: Capstone Project
- Students work on a capstone project applying concepts learned throughout the course
- Project proposal, development, and refinement stages
- Guidance and support provided by instructors and mentors

15: Project Showcase and Conclusion
- Students present their capstone projects to the class
- Peer feedback and evaluation
- Recap of key learnings and discussion on further opportunities in Databricks and PySpark