Introduction to Spark

* What is Spark?
* Spark, BigData Cluster Overview
* Setting up local environment

Spark Basics

* Resilient Distributed Datasets (RDDs)
* Spark Context
* Spark Ecosystem
* In-Memory Computations in Spark
* Using Spark Shell

Working with RDDs (Low Level API)

* Creating, Loading and Saving RDD
* Transformations on RDD
  + Map, FlatMap, Filter, Distinct, Sampling, MapValues,
  + FlatMapValues, Union, GroupBy, Cogroup, reduceByKey / groupByKey, AggregateByKey, combineByKey,
  + JOIN, MapPartition
* Actions on RDD
  + Take, Top, Collect, Take Ordered, Count, CountByValue, First,
  + Max, Min, Reduce, Aggregate, CountByKey, Lookup, CollectAsMap
* Shared Variables: Broadcast Variables & Accumulators

Job Clusters on Azure Databricks

* Job creation process in Databricks
* UI walkthrough
* Parameter Initialization
* Job Creation
* Job Scheduler
* Job Execution
* Cluster Planning

Spark SQL

* Spark SQL Architecture
* SparkSession in Spark SQL
* Working with DataFrames & DataSets
* Loading and Saving Data
  + Integrating Spark SQL with JDBC Sources (MySQL)
  + Handling CSV, JSON, XML, Avro, ORC, Parquet File Formats

Structured Streaming

* Structured Streaming Overview,Stateless & Stateful operations
* Output Modes,Window API,Event Time, Late Events, Watermark

Performance Tuning and Debugging

* Common Performance Issues & Tuning Tips
* Monitoring Driver and Executor Logs
* Custom Project Level Logging Techniques
* RePartitioning, Coalesce & Persistence Considerations
* Local & Remote Debugging of Code
* Tungsten, CO & AQE,Speculative execution
* Stragglers Identification & Handling

Spark Preformation and optimization

* Techniques can process and analyze large datasets very efficiently
* optimize Spark parameters
* pyspark ↔ pandas