**Advanced Spark**

Pre-requisite: Good knowledge in Java

Duration : 3 days

Advanced Spark Sql

* Joins
* Inner Join
* Left Outer and Right Outer joins
* Full Outer join
* Aggregations

Spark Streaming

* Introduction to SparkStreaming
* DStreams
* Working with unstructured data
* Working with Text Stream
* Using kafka streams
* Stateful Streams
* Windowing

Spark Structured Streaming

* Introduction
* Working with csv file
* Working with json file
* Checkpoint
* Different output modes
* Building ETL pipelines with Kafka
* Watermark

Deployment

* Introduction to Spark Cluster
* Standalone cluster
* Yarn cluster
* Master configurations
* Creating a cluster with 3 nodes
* Starting the Master
* Starting the slaves
* Running the Spark application in the cluster

Best Practices

* Best practices while using Dataset/DataFrame
* Catalyst Optimizer
* Coalesce over repartition
* mapPartitions over map
* Persisting and Caching Data in Memory
* Reducing Shuffles
* Optimizing Logging

Performance Tuning

* Tuning processing of Semi Structured Files
* Tuning processing of Large Structured Files
* Skewed Data
* Avoiding Spill
* Adaptive Query
* Adaptive Query Execution
* Explain Plan
* Dynamic Partition Pruning
* DataFrame functions vs UDF
* Lineage
* Fine Tuning Number of Executor and Executor Memory size
* Fine Tuning other JVM parameters