ADVANCED APACHE SPARK (5 Days)

By Dr. Vishwanath Rao

Day 1

Foundation to Spark
Spark Shell
Basic operations on Shell
Spark Context and Spark Properties Persistence in Spark
HDFS data from Spark
Working with Resilient Distributed DataSets (RDD)
Understanding RDD Loading data into RDD
Scala RDD, Paired RDD, Double RDD & General RDD Functions

Transformations, Actions and Shared Variables
Spark Operations
Spark Streaming
Introduction to Spark Streaming . Spark Structured Streaming Windowing

Delta Lakes
Streaming Architectures, Lambda Architecture
Differentiating discretized and structured streaming Linking Input Sources
Streaming Context
Discretized Streams (DStreams)
Input DStreams
Stateless Transformations on DStreams Stateful Transformations

Output Operations
Checkpointing
Caching and Persisting
Tuning and Debugging

Spark Driver (Master Process)
Spark Cluster Manager
Executors (Slave Processes)

RDD (Resilient Distributed Datasets)
DAG (Directed Acyclic Graph)

Day 2

Distributed processing using partitions efficiently
Mistakes to avoid while Optimising Apache Spark reduceByKey or groupByKey
Maintain the required size of the shuffle blocks File Formats and Delimiters
Small Data Files
No Monitoring of Job Stages
ByKey, repartition or any other operations which trigger shuffles
Reinforcement Learning

Shuffle Partitions
Spark RDD Shuffle
Spark Default Partitions
Partition Size
Adaptive Query Execution (AQE) framework
Dynamically coalescing shuffle partitions
Dynamically switching join strategies
Dynamically optimizing skew joins

Catalyst Optimizer
Custom RDD
SQL Query
DataFrame
Unresolved Logical Plan Catalog
Logical Plan

Day 3

Optimized Logical Plan Physical Plans
Cost Model
Usage of RDD
Performance Optimization Techniques

Spark SQL

Introduction to Spark SQL
Architecture and flow Querying Files as Tables
Spark SQL Overview
Import CSV
Schema Inference
Data Query Select
DataFrame.Reader DataFrame.Writer Import JSON
Data Query INNER JOINs
Data Query INNER JOINs
Group By, Order By, Window Functions
Data Query OUTER JOINs, SEMI JOIN
Custom UDF (User Defined Function) API or SQL?
Grouping,Joins,Aggregations Text file Format
JSON file Format
Hive and Spark SQL Architecture

Day 4

Apache Spark Optimization Factors and Techniques
Using Accumulators
Hive Bucketing Performance
Predicate Pushdown Optimization
Zero Data Serialization/Deserialization using Apache Arrow

Garbage Collection Tuning using G1GC Collection
Memory Management and Tuning
Data Locality
Using Collocated Joins
Caching in Spark
Executor Size
Spark Windowing Function
Watermarks Technique
Data Serialization
Serialisation
API Selection
Advance Variable
Cache and Persist
ByKey Operation
File Format Selection
Garbage Collection Tuning
Levels of Parallism - Repartition and Coalesce

Day 5

Spark Graphx
Introduction to Spark GraphX
Graph creation examples
Graph Operators Overview, Information about a Graph
Information about a graph example
Transform Graph Items
Transform graph items examples
Modify Graph Structure
Graph Neighborhood Aggregations
Neighborhood Aggregations Examples
Graph Algorithms
Joining Strategies Communication between nodes In node communication
Per node communication strategy Different Join strategies
Sort merge joins
Broadcast joins
Shuffledhash join