€TRAINING

Apache Spark Application Performance Tuning

Apache Spark Application Performance Tuning

Introduction:

This training program provides participants with essential knowledge and skills in tuning the performance of Apache Spark applications. It empowers them to optimize and troubleshoot Spark applications for improved efficiency and performance in large-scale data processing.

Program Objectives:

At the end of this program, participants will be able to:

- Understand the architecture and internals of Apache Spark.
- Identify and analyze performance bottlenecks in Spark applications.
- Apply best practices for optimizing Spark application performance.
- Utilize advanced techniques for tuning Spark jobs.
- Monitor and troubleshoot Spark applications effectively.

Targeted Audience:

- Data Engineers.
- Data Scientists.
- Big Data Developers.
- System Administrators.
- Performance Engineers.

Program Outline:

Unit 1:

Introduction to Apache Spark:

- Overview of Apache Spark architecture and components.
- Understanding Spark RDDs, DataFrames, and Datasets.
- Spark execution model and job scheduling.

- Key performance considerations in Spark applications.
- Case studies of Spark performance tuning.

Unit 2:

Identifying Performance Bottlenecks:

- Profiling and monitoring Spark applications.
- Understanding and interpreting Spark logs and metrics.
- Identifying common performance bottlenecks.
- Tools for performance analysis Spark UI, Ganglia, etc..
- Practical exercises in performance diagnostics.

Unit 3:

Optimization Techniques:

- Best practices for Spark application optimization.
- Memory management and garbage collection tuning.
- Efficient use of Spark RDDs, DataFrames, and Datasets.
- Partitioning and data locality optimization.
- Advanced join strategies and broadcast variables.

Unit 4:

Advanced Performance Tuning:

- Tuning Spark SQL and Catalyst optimizer.
- Optimizing shuffle operations and spill behavior.
- Configuring and tuning Spark's cluster manager YARN, Mesos, etc..
- Parameter tuning for Spark jobs and executors.
- Case studies of advanced performance tuning.

Unit 5:

Monitoring and Troubleshooting:

- Continuous monitoring of Spark applications.
- Setting up and using monitoring tools Prometheus, Grafana, etc..
- Troubleshooting common Spark performance issues.
- Debugging and resolving Spark job failures.
- Best practices for maintaining Spark application performance.