APACHE SPARK (5 DAYS)

By. Dr. Vishwanath Rao

Introduction to Functional Programming

- What is Functional Programming (FP)?
- Terminology: First-Class and Higher-Order Functions
- Terminology: Lambda vs Closure
- A Short List of Languages that Support FP
- FP with Java
- FP With JavaScript
- Imperative Programming in JavaScript
- The JavaScript map (FP) Example
- The JavaScript reduce (FP) Example
- Using reduce to Flatten an Array of Arrays (FP) Example
- The JavaScript filter (FP) Example
- Common High-Order Functions in Python
- Common High-Order Functions in Scala
- Elements of FP in R
- Summary

Introduction to Apache Spark

- What is Spark
- A Short History of Spark
- Where to Get Spark?
- The Spark Platform
- Spark Logo
- Common Spark Use Cases
- Languages Supported by Spark
- Running Spark on a Cluster
- The Driver Process
- Spark Applications
- Spark Shell
- The spark-submit Tool
- The spark-submit Tool Configuration
- The Executor and Worker Processes
- The Spark Application Architecture
- Interfaces with Data Storage Systems
- Limitations of Hadoop's MapReduce
- Spark vs MapReduce
- Spark as an Alternative to Apache Tez
- The Resilient Distributed Dataset (RDD)
- Spark Streaming (Micro-batching)
- Spark SQL
- Example of Spark SQL
- Spark Machine Learning Library

- GraphX
- Spark vs R
- Summary

Hadoop Distributed File System Overview

- Hadoop Distributed File System (HDFS)
- HDFS High Availability
- HDFS "Fine Print"
- Storing Raw Data in HDFS
- Hadoop Security
- HDFS Rack-awareness
- Data Blocks
- Data Block Replication Example
- HDFS NameNode Directory Diagram
- Accessing HDFS
- Examples of HDFS Commands
- Other Supported File Systems
- WebHDFS
- Examples of WebHDFS Calls
- Client Interactions with HDFS for the Read Operation
- Read Operation Sequence Diagram
- Client Interactions with HDFS for the Write Operation
- Communication inside HDFS
- Summary

The Spark Shell

- The Spark Shell
- The Spark Shell UI
- Spark Shell Options
- Getting Help
- The Spark Context (sc) and SQL Context (sqlContext)
- The Shell Spark Context
- Loading Files
- Saving Files
- Basic Spark ETL Operations
- Summary

Spark RDDs

- The Resilient Distributed Dataset (RDD)
- Ways to Create an RDD
- Custom RDDs
- Supported Data Types
- RDD Operations
- RDDs are Immutable
- Spark Actions

- RDD Transformations
- Other RDD Operations
- Chaining RDD Operations
- RDD Lineage
- The Big Picture
- What May Go Wrong
- Checkpointing RDDs
- Local Checkpointing
- Parallelized Collections
- More on parallelize() Method
- The Pair RDD
- Where do I use Pair RDDs?
- Example of Creating a Pair RDD with Map
- Example of Creating a Pair RDD with keyBy
- Miscellaneous Pair RDD Operations
- RDD Caching
- RDD Persistence
- The Tachyon Storage
- Summary

Shared Variables in Spark

- Shared Variables in Spark
- Broadcast Variables
- Creating and Using Broadcast Variables
- Example of Using Broadcast Variables
- Accumulators
- Creating and Using Accumulators
- Example of Using Accumulators
- Custom Accumulators
- Summary

Parallel Data Processing with Spark

- Running Spark on a Cluster
- Spark Stand-alone Option
- The High-Level Execution Flow in Stand-alone Spark Cluster
- Data Partitioning
- Data Partitioning Diagram
- Single Local File System RDD Partitioning
- Multiple File RDD Partitioning
- Special Cases for Small-sized Files
- Parallel Data Processing of Partitions
- Spark Application, Jobs, and Tasks
- Stages and Shuffles
- The "Big Picture"
- Summary

Introduction to Spark SQL

- What is Spark SQL?
- Uniform Data Access with Spark SQL
- Hive Integration
- Hive Interface
- Integration with BI Tools
- Spark SQL is No Longer Experimental Developer API!
- What is a DataFrame?
- The SQLContext Object
- The SQLContext API
- Changes Between Spark SQL 1.3 to 1.4
- Example of Spark SQL (Scala Example)
- Example of Working with a JSON File
- Example of Working with a Parquet File
- Using JDBC Sources
- JDBC Connection Example
- Performance & Scalability of Spark SQL
- Summary

Graph Processing with GraphX

- What is GraphX?
- Supported Languages
- Vertices and Edges
- Graph Terminology
- Example of Property Graph
- The GraphX API
- The GraphX Views
- The Triplet View
- Graph Algorithms
- Graphs and RDDs
- Constructing Graphs
- Graph Operators
- Example of Using GraphX Operators
- GraphX Performance Optimization
- The PageRank Algorithm
- GraphX Support for PageRank
- Summary

Machine Learning Algorithms

- Supervised vs Unsupervised Machine Learning
- Supervised Machine Learning Algorithms
- Unsupervised Machine Learning Algorithms
- Choose the Right Algorithm
- Life-cycles of Machine Learning Development
- Classifying with k-Nearest Neighbors (SL)

- k-Nearest Neighbors Algorithm
- k-Nearest Neighbors Algorithm
- The Error Rate
- Decision Trees (SL)
- Random Forests
- Unsupervised Learning Type: Clustering
- K-Means Clustering (UL)
- K-Means Clustering in a Nutshell
- Regression Analysis
- Logistic Regression
- Summary

The Spark Machine Learning Library

- What is MLlib?
- Supported Languages
- MLlib Packages
- Dense and Sparse Vectors
- Labeled Point
- Python Example of Using the LabeledPoint Class
- LIBSVM format
- An Example of a LIBSVM File
- Loading LIBSVM Files
- Local Matrices
- Example of Creating Matrices in MLlib
- Distributed Matrices
- Example of Using a Distributed Matrix
- Classification and Regression Algorithm
- Clustering
- Summary

Spark Streaming

- What is Spark Streaming?
- Spark Streaming as Micro-batching
- Use Cases
- Some "Competition"
- Spark Streaming Features
- How It Works
- Basic Data Stream Sources
- Advanced Data Stream Sources
- The DStream Object
- DStream - RDD Diagram
- The Operational DStream API
- DStream Output Operations
- The
- StreamingContext Object
- TCP Text Streams Example (in Scala)

- Accessing the Underlying RDDs
- The Sliding Window Concept
- The Sliding Window Diagram
- The Window Operations
- A Windowed Computation Example (Scala)
- Points to Remember
- Other Points to Remember
- Summary