Apache Spark (By Ernesto Lee)

Overview

Introduction to Apache Spark

Description

Apache Spark is an open-source distributed general-purpose cluster-computing framework. Spark provides an interface for programming entire clusters with implicit data parallelism and fault tolerance. It is used for processing and analyzing a large amount of data. Just like Hadoop MapReduce, it also works with the system to distribute data across the cluster and process the data in parallel.

Labs

Labs for this course are available at path shared below. Elev8ed Labs (powered by Jupyter) will be accessible at the port given to you by your instructor.

1. Apache Spark Scala Basics
2. Apache Spark Scala Data Types and Loops
3. Apache Spark Scala Advanced
4. Apache Spark Installation
5. Apache Spark Creating RDDs from Spark-Shell
6. Apache Spark WordCount
7. Apache Spark WebUI
8. Apache Spark RDD Caching and Persistence
9. Apache Spark Paired RDD
10. Apache Spark Paired RDD Advanced
11. Apache Spark Paired RDD Joins & Actions
12. Apache Spark Accumulators V1
13. Apache Spark Accumulators V2
14. Apache Spark Accumulators Custom
15. Apache Spark Broadcast Variables
16. Apache Spark - Creating Data Frame using Data Source API
17. Apache Spark - Creating Data Frame from an RDD and StructType
18. Apache Spark - Querying data using Spark SQL
19. Apache Spark - Joins using Spark SQL
20. Apache Spark - Creating Dataset using Data Source API
21. Apache Spark - Creating Dataset from an RDD
22. Apache Spark - Aggregate and Collection Functions
23. Apache Spark Date/Time Functions
24. Apache Spark Math and String Functions
25. Apache Spark Window Functions
26. Apache Spark Currying and Partially Applied Functions
27. Apache Spark Writing User Defined Function
28. Apache Spark Writing Untyped UDAF
29. Apache Spark Typed UDAF
30. Apache Spark File Formats - Text
31. Apache Spark File Formats - CSV and JSON
32. Apache Spark File Formats - Parquet and ORC
33. Apache Spark File Formats - Hadoop and Sequence