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**

   ```
   * <host-ip>:<port>/lab/workspaces/lab1
   ```

2. **Apache Spark Scala Data Types and Loops**

   ```
   * <host-ip>:<port>/lab/workspaces/lab2
   ```

3. **Apache Spark Scala Advanced**

   ```
   * <host-ip>:<port>/lab/workspaces/lab3
   ```

4. **Apache Spark Installation**

   ```
   * <host-ip>:<port>/lab/workspaces/lab4
   ```

5. **Apache Spark Creating RDDs from Spark-Shell**

   ```
   * <host-ip>:<port>/lab/workspaces/lab5
   ```

6. **Apache Spark WordCount**

   ```
   * <host-ip>:<port>/lab/workspaces/lab6
   ```

7. **Apache Spark WebUI**

   ```
   * <host-ip>:<port>/lab/workspaces/lab7
   ```

8. **Apache Spark RDD Caching and Persistence**

   ```
   * <host-ip>:<port>/lab/workspaces/lab8
   ```

9. **Apache Spark Paired RDD**

   ```
   * <host-ip>:<port>/lab/workspaces/lab9
   ```

10. **Apache Spark Paired RDD Advanced**

```
* <host-ip>:<port>/lab/workspaces/lab10
```

11. **Apache Spark Paired RDD Joins & Actions**

    ```
    * <host-ip>:<port>/lab/workspaces/lab11
    ```

12. **Apache Spark Accumulators V1**

    ```
    * <host-ip>:<port>/lab/workspaces/lab12
    ```

13. **Apache Spark Accumulators V2**

    ```
    * <host-ip>:<port>/lab/workspaces/lab13
    ```

14. **Apache Spark Accumulators Custom**

    ```
    * <host-ip>:<port>/lab/workspaces/lab14
    ```

15. **Apache Spark Broadcast Variables**

    ```
    * <host-ip>:<port>/lab/workspaces/lab15
    ```

16. **Apache Spark - Creating Data Frame using Data Source API**

    ```
    * <host-ip>:<port>/lab/workspaces/lab16
    ```

17. **Apache Spark - Creating Data Frame from an RDD and StructType**

    ```
    * <host-ip>:<port>/lab/workspaces/lab17
    ```

18. **Apache Spark - Querying data using Spark SQL**

    ```
    * <host-ip>:<port>/lab/workspaces/lab18
    ```

19. **Apache Spark - Joins using Spark SQL**

    ```
    * <host-ip>:<port>/lab/workspaces/lab19
    ```

20. **Apache Spark - Creating Dataset using Data Source API**

    ```
    * <host-ip>:<port>/lab/workspaces/lab20
    ```

21. **Apache Spark - Creating Dataset from an RDD**

    ```
    * <host-ip>:<port>/lab/workspaces/lab21
    ```

22. **Apache Spark - Aggregate and Collection Functions**

    ```
    * <host-ip>:<port>/lab/workspaces/lab22
    ```

23. **Apache Spark Date/Time Functions**

    ```
    * <host-ip>:<port>/lab/workspaces/lab23
    ```

24. **Apache Spark Math and String Functions**

```
* <host-ip>:<port>/lab/workspaces/lab24
```

25. **Apache Spark Window Functions**

```
* <host-ip>:<port>/lab/workspaces/lab25
```

26. **Apache Spark Currying and Partially Applied Functions**

```
* <host-ip>:<port>/lab/workspaces/lab26
```

27. **Apache Spark Writing User Defined Function**

```
* <host-ip>:<port>/lab/workspaces/lab27
```

28. **Apache Spark Writing Untyped UDAF**

```
* <host-ip>:<port>/lab/workspaces/lab28
```

29. **Apache Spark Typed UDAF**

```
* <host-ip>:<port>/lab/workspaces/lab29
```

30. **Apache Spark File Formats - Text**

```
* <host-ip>:<port>/lab/workspaces/lab30
```

31. **Apache Spark File Formats - CSV and JSON**

```
* <host-ip>:<port>/lab/workspaces/lab31
```

32. **Apache Spark File Formats - Parquet and ORC**

```
* <host-ip>:<port>/lab/workspaces/lab32
```

33. **Apache Spark File Formats - Hadoop and Sequence**

```
* <host-ip>:<port>/lab/workspaces/lab33
```