Apache Spark

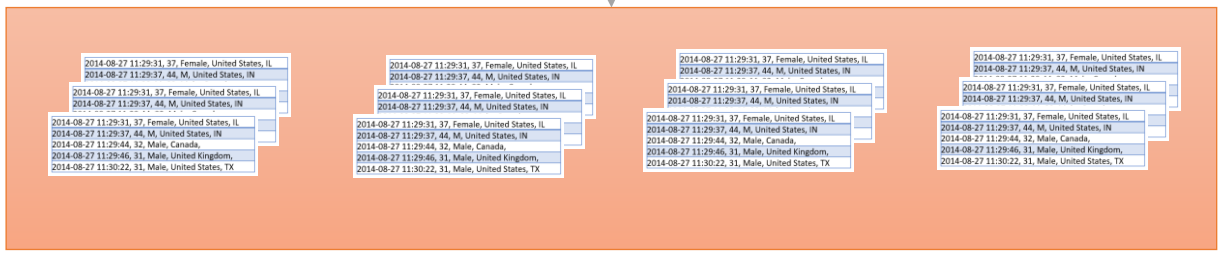
Spark RDD APIs

RDD – Resilient Distributed Dataset

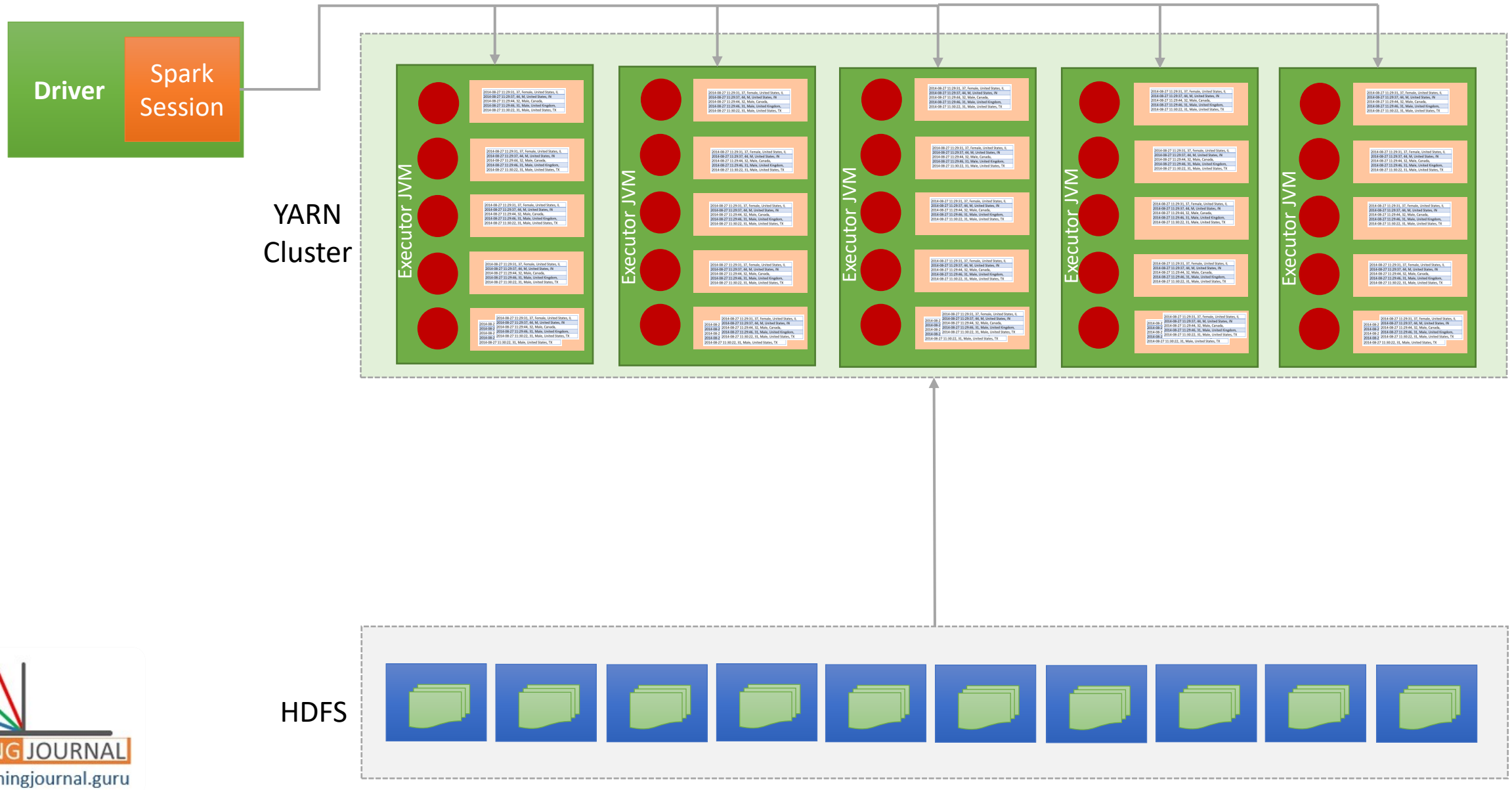
Dataset –

2014-08-27 11:29:31, 37, Female, United States, IL
2014-08-27 11:29:37, 44, M, United States, IN
2014-08-27 11:29:44, 32, Male, Canada,
2014-08-27 11:29:46, 31, Male, United Kingdom,
2014-08-27 11:30:22, 31, Male, United States, TX

Distributed –



Spark RDD APIs



LEARNING JOURNAL
www.learningjournal.guru

Creating Data frame

1. Create a SparkConf object
2. Create a Spark Session using your SparkConf
3. Use the spark session to read the data file

```
val sparkAppConf = new SparkConf
sparkAppConf.setAppName("Hello Spark").setMaster("local[3]")
val spark = SparkSession.builder().config(sparkAppConf).getOrCreate()
val surveyRawDF = spark.read.option("header", "true").option("inferSchema", "true").csv(args(0))
```


LEARNING JOURNAL
www.learningjournal.guru