Spark DataFrame

Why DataFrame?

- 3 Data Structures of Spark
  - RDD
  - DataFrame
  - Datasets
- DataFrame definition From Spark documentation

"A DataFrame is a *Dataset* organized into named columns. It is conceptually equivalent to a table in a relational database or a data frame in R/Python, but with richer optimizations under the hood"

Key Features Of DataFrame

- Automatic Optimization of code.
- you can run SQL queries on DataFrame using spark SQL
- Language support available for pyspark, scala, R and java
- Lower Learning Curve
- Provide DataSource API to read DataFrame Multiple Formats

Create a DataFrame

- From a File directly

  val df = spark.read.json("/path/of/file.json")

- From a RDD

  val data =Seq(…)

  val rdd = sc.makeRDD(data).map(x => Row(x._1,x._2,…))

  import org.apache.spark.sql.types._
  val schema = StructType(Seq(StructField(…))

  val df = spark.createDataFrame(rdd,schema)