# 连接Spark

## 参考网址: <http://spark.apache.org/docs/2.3.1/rdd-programming-guide.html>

## 构建spark(以2.3.1为例)

### 引入Maven包

|  |
| --- |
| <**dependency**>  <**groupId**>org.apache.spark</**groupId**>  <**artifactId**>spark-core\_2.11</**artifactId**>  <**version**>2.3.1</**version**> </**dependency**> |

### 导类

|  |
| --- |
| **import** **org.apache.spark.SparkContext**  **import** **org.apache.spark.SparkConf** |

## 初始化Spark

每个JVM只能激活一个SparkContext, 创建新的SparkContex前要stop()前面的SparkContext.

|  |
| --- |
| **val** conf **=** **new** **SparkConf**().setAppName(appName).setMaster(master)  **new** **SparkContext**(conf) |

## 连接HDFS集群(额外介绍)

### 引入Maven包

|  |
| --- |
| <**dependency**>  <**groupId**>org.apache.hadoop</**groupId**>  <**artifactId**>hadoop-client</**artifactId**>  <**version**><your-hdfs-version></**version**> </**dependency**> |

# RDD创建

## 通过SparkContext创建

|  |
| --- |
| **val** data **=** **Array**(1, 2, 3, 4, 5)  **val** distData **=** sc.parallelize(data) |

### Note: 指定分区

|  |
| --- |
| sc.parallelize(data, 10)) |

## 通过Hadoop支持的任意存储源(本地文件系统、HDFS、Hbase、Hive等)

|  |
| --- |
| scala> **val** distFile **=** sc.textFile("data.txt")  distFile**:** org.apache.spark.rdd.RDD[String] **=** data.txt **MapPartitionsRDD**[10] at textFile at <console**>:**26 |

#### Note: HDFS文件: hdfs://

## 操作RDD

# RDD转化为DataFrame

参考网址: <https://blog.csdn.net/weixin_39216383/article/details/80494630>

<https://blog.csdn.net/lvtula/article/details/93851572>

## 通过反射的方式

|  |
| --- |
| 通过反射的方式, 需要事先知道这份数据的Schema信息. 应用场景: 读取HDFS数据等. |

## 通过编码实现

|  |
| --- |
| 主要调用spark.createDataFrame(rdd, schema). 其中rdd是Seq(Row)类型, schema是StructType类型 |