Spark高级练习一

**练习1：**分析整数数据集

1. 创建com.aura.spark.BasicPracticeOne类，定义JavaSparkContext，输入下列代码

|  |
| --- |
| package com.aura.spark;  import org.apache.spark.SparkConf; import org.apache.spark.api.java.JavaRDD; import org.apache.spark.api.java.JavaSparkContext;  import java.util.ArrayList; import java.util.List;  public class BasicPracticeOne {  public static void main(String[] args) {  SparkConf conf = new SparkConf()  .setMaster("local[1]")  .setAppName("basicPracticeOne");  JavaSparkContext jsc = new JavaSparkContext(conf);   List<Integer> result = new ArrayList();  for(int i = 100; i <= 1000; i++) result.add(i);  JavaRDD<Integer> input = jsc.parallelize(result, 7);   // TODO add your code here  jsc.stop();  } } |

2. 在//TODO注释处添加加入代码, 完成下列功能：

* 打印input中前5个数据
* 输出input中所有元素和
* 输出input中所有元素的平均值
* 统计input中偶数的个数，并打印前5个

3. 在IDE中运行代码检验结果

|  |
| --- |
| head 5 elements are:[100, 101, 102, 103, 104]  sum is:495550  avg is:550.0  count of even number is:451  head 5 of even number is:[100, 102, 104, 106, 108] |

**练习2**：高级RDD算子

1. 创建com.aura.spark.BasicPracticeTwo类，定义JavaSparkContext，输入下列代码

|  |
| --- |
| package com.aura.spark;  import org.apache.spark.SparkConf; import org.apache.spark.api.java.JavaRDD; import org.apache.spark.api.java.JavaSparkContext;  import java.util.ArrayList; import java.util.Arrays; import java.util.List;  public class BasicPracticeTwo {  public static void main(String[] args) {  SparkConf conf = new SparkConf()  .setMaster("local[1]")  .setAppName("basicPracticeTwo");  JavaSparkContext jsc = new JavaSparkContext(conf);   List<Integer> data = Arrays.asList(1,2,3,4,5, 6);  JavaRDD<Integer> rdd1 = jsc.parallelize(data, 3);  List<Integer> data2 =Arrays.asList(7,8,9,10,11);  JavaRDD<Integer> rdd2 = jsc.parallelize(data, 2);  List<Integer> data3=Arrays.asList(12,13,14,15,16, 17, 18, 19, 20, 21);  JavaRDD<Integer> rdd3 = jsc.parallelize(data3, 3);   // TODO add your code here  jsc.stop();  } } |

2. 在//TODO注释处添加加入代码, 完成下列功能：

* 使用union连接rdd1和rdd2，生成rdd4
* 使用glom打印rdd4的各个partition
* 使用coalesce将rdd4的分区数改为3，并生成rdd5
* 使用repartition将rdd4的分区数改为10，并生成rdd6
* 使用glom分别打印rdd5和rdd6中的partition元素均匀性

3. 在IDE中运行代码，检查结果

|  |
| --- |
| rdd4 has 5 partitions  Partition 0 is:[1, 2]  Partition 1 is:[3, 4]  Partition 2 is:[5, 6]  Partition 3 is:[1, 2, 3]  Partition 4 is:[4, 5, 6]  rdd5 has 3 partitions  Partition 0 is:[1, 2]  Partition 1 is:[3, 4, 5, 6]  Partition 2 is:[1, 2, 3, 4, 5, 6]  rdd6 has 10 partitions  Partition 0 is:[2]  Partition 1 is:[1, 3]  Partition 2 is:[2, 4]  Partition 3 is:[5]  Partition 4 is:[6]  Partition 5 is:[]  Partition 6 is:[3]  Partition 7 is:[4]  Partition 8 is:[5]  Partition 9 is:[6, 1] |

**练习3**：Key/Value RDD算子

1. 创建com.aura.spark.BasicPracticeThree类，定义JavaSparkContext，输入下列代码

|  |
| --- |
| package com.aura.spark;  import org.apache.spark.SparkConf; import org.apache.spark.api.java.JavaPairRDD; import org.apache.spark.api.java.JavaSparkContext; import scala.Tuple2;  import java.util.Arrays; import java.util.List;  public class BasicPracticeThree {  public static void main(String[] args) {  SparkConf conf = new SparkConf()  .setMaster("local[1]")  .setAppName("basicPracticeThree");  JavaSparkContext jsc = new JavaSparkContext(conf);   jsc.setLogLevel("error");  List<Tuple2<String, Integer>> data = Arrays.asList(  new Tuple2("coffee", 1),  new Tuple2("coffee", 3),  new Tuple2("panda", 4),  new Tuple2("coffee", 5),  new Tuple2("street", 2),  new Tuple2("panda", 5)  );  JavaPairRDD<String, Integer> input = jsc.parallelizePairs(data);   // TODO add your code here  jsc.stop();  } }  class SumAndCount implements Serializable {  final int sum;  final int count;  public SumAndCount(int sum, int count) {  this.sum = sum;  this.count = count;  }  public double average() {  if (count != 0) {  return sum\*1.0/count;  } else {  return 0.0;  }  } } |

2. 在注释//TODO处编写代码完成下列功能

计算相同Key对应的的所有value的平均值，并输出到目录/tmp/output下

3. 查看/tmp/output下的结果

|  |
| --- |
| (coffee,3.0)  (street,2.0)  (panda,4.5) |

**练习4**：沃尔玛交易流水分析

1. 在/tmp目录下创建input.txt文件，输入下列数据，分别表示分店ID（id），交易量（x），交易额（y）和利润（z），类型均为整数

|  |
| --- |
| 1,23,5600,5  2,30,5800,7  1,27,5000,10  3,24,6900,5  2,45,5800,7  2,28,5800,7 |

2. 创建com.aura.spark.BasicPracticeFour类，定义JavaSparkContext，输入下列代码

|  |
| --- |
| package com.aura.spark;  import org.apache.spark.SparkConf; import org.apache.spark.api.java.JavaSparkContext;  public class BasicPracticeFour {  public static void main(String[] args) {  SparkConf conf = new SparkConf()  .setMaster("local[1]")  .setAppName("basicPracticeFour");  JavaSparkContext jsc = new JavaSparkContext(conf);    // TODO add your code here  jsc.stop();  } } |

3. 在注释//TODO处添加代码，完成下列过程：

* 从/tmp/input.txt中读取数据到RDD中，
* 用RDD的transformation函数实现下列功能SELECT id, SUM(x), MAX(y), MIN(z), AVERAGE(x) FROM T GROUP BY id;
* 将结果RDD输出到控制台中

4. 检查结果

|  |
| --- |
| ID 1, max(y)=5600, min(z)=5, avg(x)=25.000000  ID 3, max(y)=6900, min(z)=5, avg(x)=24.000000  ID 2, max(y)=5800, min(z)=7, avg(x)=34.333333 |