大数据开发技术

东北林业大学

卢洋

第三章

MapReduce框架原理

自定义OutputFormat

OutputFormat是MapReduce输出的基类，所有实现MapReduce输出都实现了OutputFormat接口。下面介绍几种常见的OutputFormat实现类。

1. 文本输出TextOutputFormat
   默认的输出格式是TextOutputFormat，它把每条记录写为文本行。它的键和值可以是任意类型，因为TextOutputFormat调用toString()方法把它们转换为字符串。

2. SequenceFileOutputFormat
   将SequenceFileOutputFormat输出作为后续MapReduce任务的输入，这便是一种好的输出格式，因其格式紧凑，很容易被压缩。

3. 自定义OutputFormat
   根据用户需求，自定义实现输出。

自定义OutputFormat使用场景及步骤

1. 使用场景

   为了实现控制最终文件的输出路径和输出格式，可以自定义OutputFormat。

   例如：要在一个MapReduce程序中根据数据的不同输出两类结果到不同目录，这类灵活的输出需求可以通过自定义OutputFormat来实现。

2. 自定义OutputFormat步骤

(1) 自定义一个类继承FileOutputFormat;

(2) 改写RecordWriter，具体改写输出数据的方法write().