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InputFormat

自定义InputFormat

- 在实际开发中，Hadoop自带的InputFormat类型并不能满足所有应用场景，需要自定义InputFormat来解决实际问题。

自定义InputFormat步骤

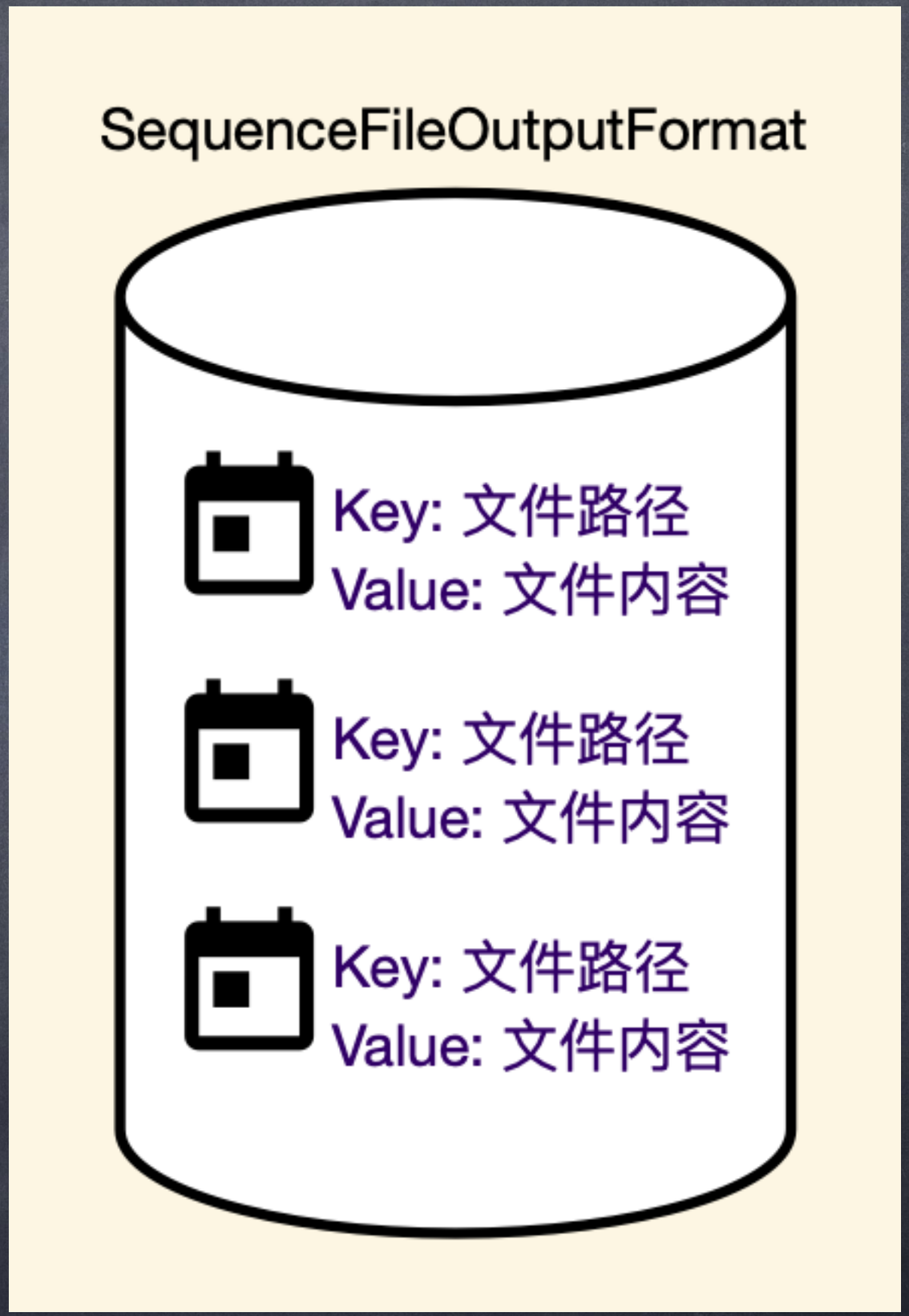
1. 自定义一个类，继承FileInputFormat;

2. 改写RecordReader，实现一次读取一个完整文件封装为KV;

3. 在输出是使用SequenceFileOutputFormat输出合并文件.

处理小文件，最后输出大文件

SequenceFileOutputFormat



处理小文件

1. HAR;

归档

2. CombineTextInputFormat;

3. 自定义InputFormat.