

# 构思思路

## FileInputFormat

由于Hadoop无法直接对.xlsx文件做处理，所以我们需要对数据做预处理。有两种方法：

- 1、自定义InputFormat类与RecordReader类，实现对excel文件的split，输出split文件键值对。
- 2、将excel文件提前转为.csv文件，这样可以用hadoop里的默认类处理

由于第一种方法需要org.apache.poi库来帮助我们读.xlsx文件，而这个库会有冲突，导致hadoop任务无法正常进行（我也不知道为什么，我没有很好的解决方法）

所以我们采用第二种，这样我们可以直接利用Hadoop内置的KeyValueTextInputFormat来获取split的键值对

## Mapper

mapper的实现很简单，因为我们使用的是KeyValueTextInputFormat，所以传入的是什么键值对直接就输出就ok。


## Reducer

reducer的任务就是去重，对于传入的(key,Iterable values),我们利用java的Set类实现去重操作，然后对于Set里面的每一个值，我们都将其输出。需要注意的是，Text类应该没有实现HashCode()与equal()方法，所以Set的元素类型如果是Text的话可能会由于Text对象无法确定是否相等，导致最终的运行结果会有错误。(实践出来的)

# 运行结果


ID	User	Name	Application Type	Application Tags	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU Vcores	Allocated Memory MB	Allocated GPUs	Reserve CPU Vcores
1698156979633_0003	lrz	homework_5	MAPREDUCE		default	0	Tue Oct 24 22:59:12 +0800 2023	Tue Oct 24 22:59:15 +0800 2023	Tue Oct 24 23:00:50 +0800 2023	FINISHED	SUCCEEDED	1	1	2048	N/A	0


Open ▾





part-r-00000

Save









~/financial\_data/homework/homework\_5/output

1 101 AAPL

2 101 CSCO

3 101 KO

4 101 JPM

5 101 HON

6 101 NKE

7 101 GS

8 101 MMM

9 101 AMGN

10 101 DOW

11 101 UNH

12 101 DIS

13 101 MSFT

14 101 V

15 101 INTC

16 101 CAT

17 101 AXP

18 101 HD

19 101 CRM

20 101 BA

21 102 AHG

22 102 AAPL

23 102 CSCO

24 102 ACXP

25 102 TT00

26 102 GRTS

27 102 NFTG

28 102 MSFT

29 102 JOAN

30 102 PRZO

31 102 ADAF

32 102 WBA

33 102 OMGA

34 102 VCNX

35 102 EEIQ

36 102 ORGS

37 102 COYA

Plain Text ▾

Tab Width: 8 ▾

Ln 9, Col 13 ▾

INS