

四种常见的MapReduce设计模式

使用MapReduce解决任何问题之前，我们需要考虑如何设计。并不是任何时候都需要map和reduce job。

MapReduce设计模式(MapReduce Design Pattern)

整个MapReduce作业的阶段主要可以分为以下四种：

- 1、Input-Map-Reduce-Output
- 2、Input-Map-Output
- 3、Input-Multiple Maps-Reduce-Output
- 4、Input-Map-Combiner-Reduce-Output

下面我将一一介绍哪种场景使用哪种设计模式。

Input-Map-Reduce-Output

Input → Map → Reduce → Output

如果我们需要做一些聚合操作(aggregation)，我们就需要使用这种模式。

场景	计算各性别员工薪水平均值
Map(Key, Value)	Key: Gender Value: Their Salary
Reduce	对Gender进行Group by，并计算每种性别的总薪水

Input-Map-Output

Input → Map → Output

如果我们仅仅想改变输入数据的格式，这时候我们可以使用这种模式。

场景	对性别进行处理
Map(Key, Value)	Key : Employee Id

Value : Gender ->

if Gender is Female/ F/ f/ 0 then converted to F
else if Gender is Male/M/m/1 then convert to M

Input-Multiple Maps-Reduce-Output

Input1 → Map1 →
 Reduce → Output
Input2 → Map2 →

在这种设计模式中，我们有两个输入文件，其文件的格式都不一样，
文件一的格式是性别作为名字的前缀，比如：Ms. Shital Katkar或Mr. Krishna Katkar
文件二的格式是性别的格式是固定的，但是其位置不固定，比如 Female/Male, 0/1, F/M

场景

Map(Key, Value)

对性别进行处理

Map 1 (For input
1)

：我们需要将性别从名字中分割出来，然后根据
前缀来确定性别，然后得到
(Gender,Salary)键值对；

Map 2 (For input
2)

：这种情况程序编写比较直接，处理固定格式的
性别，然后得到(Gender,Salary)键值对。

Reduce

对Gender进行Group
by，并计算每种性别的总薪水

Input-Map-Combiner-Reduce-Output

Input → Map → Combiner → Reduce → Output

在MapReduce中，Combiner也被成为Reduce，其接收Map端的输出作为其输入，并且将输出的 key-value
键值对作为Reduce的输入。Combiner的使用目的是为了减少数据传入到Reduce的负载。

在MapReduce程序中，20%的工作是在Map阶段执行的，这个阶段也被成为数据的准备阶段，各阶段的工作是并行进行的。

80%的工作是在Reduce阶段执行的，这个阶段被成为计算阶段，其不是并行的。因此，次阶

段一般要比Map阶段要满。为了节约时间，一些在Reduce阶段处理的工作可以在combiner阶段完成。

假设我们有5个部门(departments)，我们需要计算个性别的总薪水。但是计算薪水的规则有点奇怪，比如某个性别的总薪水大于200k，那么这个性别的总薪水需要加上20k；如果某个性别的总薪水大于100k，那么这个性别的总薪水需要加上10k。如下：

Map阶段：

Dept 1: Male<10,20,25,45,15,45,25,20>,Female <10,30,20,25,35>

Dept 2: Male<15,30,40,25,45>,Female <20,35,25,35,40>

Dept 3: Male<10,20,20,40>,Female <10,30,25,70>

Dept 4: Male<45,25,20>,Female <30,20,25,35>

Dept 5: Male<10,20>,Female <10,30,20,25,35>

Combiner阶段：

Dept 1 : Male <250,20>,Female <120,10>

Dept 2 : Male <155,10>,Female <175,10>

Dept 3 : Male <90,00>,Female <135,10>

Dept 4 : Male <90,00>,Female <110,10>

Dept 5 : Male <30,00>,Female <130,10>

Reduce阶段：

Male< 250,20,155,10,90,90,30> , Female<120,10,175,10,135,10,110,10,130,10>

Output：

Male<645>,Female<720>

以上四种MapReduce模式只是最基本的，我们可以根据自己问题设计不一样的设计模式。

本文翻译自：<https://dzone.com/articles/mapreduce-design-patterns>

本博客文章除特别声明，全部都是原创！

原创文章版权归过往记忆大数据（[过往记忆](#)）所有，未经许可不得转载。

本文链接：[【】（）](#)