



大数据分析实验

MapReduce



目录

一 任务背景

二 任务描述

1. 基本任务

2. 进阶任务

三 算法流程

四 验收流程



华中科技大学
计算机科学与技术学院
School of Computer Science & Technology, HUST

一 任务背景

一 任务背景



- 1、理解map-reduce算法思想与流程;
- 2、应用map-reduce思想解决问题;
- 3、掌握并应用combine与shuffle过程。



华中科技大学
计算机科学与技术学院
School of Computer Science & Technology, HUST

二 任务描述

二 任务描述

◆ 基本任务

- 实验数据：提供9个预处理过的文件夹（folder_1-9）模拟9个分布式节点中的数据，每个源文件夹中包含大约6千个文件，每个文件标题为维基百科条目标题，内容为对应的网页内容。提供words.txt文件作为待统计的词汇，words.txt为所有文件的标题。

The screenshot displays the experimental data setup. On the left, a Windows File Explorer window shows the directory path: 此电脑 > 科研盘 (D:) > Download > 大数据分析实验 > MapReduce > source_data. A red circle highlights the subfolders folder_1 through folder_9. In the center, a Notepad++ window titled 'D:\Download\大数据分析实验\MapReduce\source_data\folder_1\aaach.txt - Notepad++' shows the content of a file named 'aaach.txt'. The text in the file includes a heading ''Aach'' can mean:', followed by two bulleted items: '* [[Aach, Baden-Württemberg]]' and '* [[Aach, Rhineland-Palatinate]]', and a disambiguation section '{{disambig}}'. On the right, another Notepad++ window titled 'words.txt' displays a list of 24 Wikipedia article titles, starting with 'Air' and ending with 'Bankruptcy'.

名称	修改日期	类型	大小
folder_1	2024/3/22 17:49	文件夹	
folder_2	2024/3/22 17:49	文件夹	
folder_3	2024/3/22 17:49	文件夹	
folder_4	2024/3/22 17:50	文件夹	
folder_5	2024/3/22 17:50	文件夹	
folder_6	2024/3/22 17:50	文件夹	
folder_7	2024/3/22 17:50	文件夹	
folder_8	2024/3/22 17:50	文件夹	
folder_9	2024/3/22 17:50	文件夹	

```
''Aach'' can mean:
* [[Aach, Baden-Württemberg]]
* [[Aach, Rhineland-Palatinate]]

{{disambig}}

{{Short pages monitor}} < !-- This long comment was added to the page to pre
```

```
1 Air
2 Arithmetic
3 Andouille
4 Farming
5 Addition
6 Albigenian
7 Abbreviation
8 Angel
9 Algebra
10 As
11 Atom
12 Astronomy
13 Asteroid
14 Anatomy
15 Austria
16 Armenia
17 Animalia
18 Application
19 Animal
20 Acceleration
21 Boot
22 Bankruptcy
23 Browser
```

◆ 基本任务

- 要求应用map-reduce思想，模拟9个map节点与3个reduce节点实现对维基百科条目词汇的词频的统计。
- map节点输出 $\langle ((\text{title1}, \text{key1}), 1), \dots, ((\text{titlem}, \text{keyn}), 1) \rangle$ ，其中key为文件title.txt中出现的且在words.txt中词汇。
- 同时，最终的reduce节点输出出现次数最多的前1000个词汇，以及这些词汇的跳转关系（作为后面实验二和实验三的输入数据）。

◆ 进阶任务

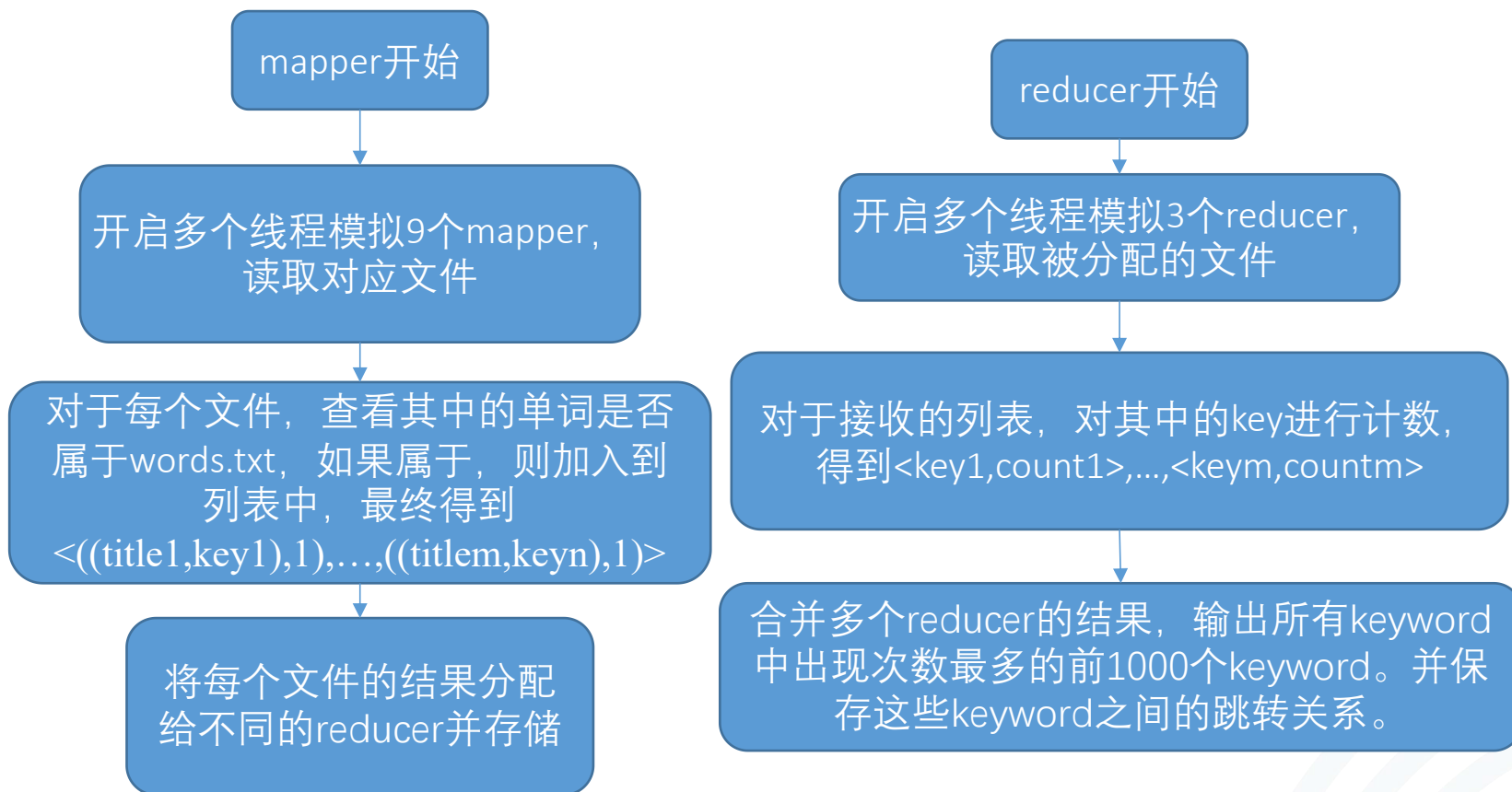
- 掌握并应用combine与shuffle过程
- 1、Shuffle过程：map节点通过shuffle过程将任务大致均分给reduce节点。
- 2、Combine过程：map节点通过combine过程压缩输出内容，减少map节点与reduce节点通信。



华中科技大学
计算机科学与技术学院
School of Computer Science & Technology, HUST

三 算法流程

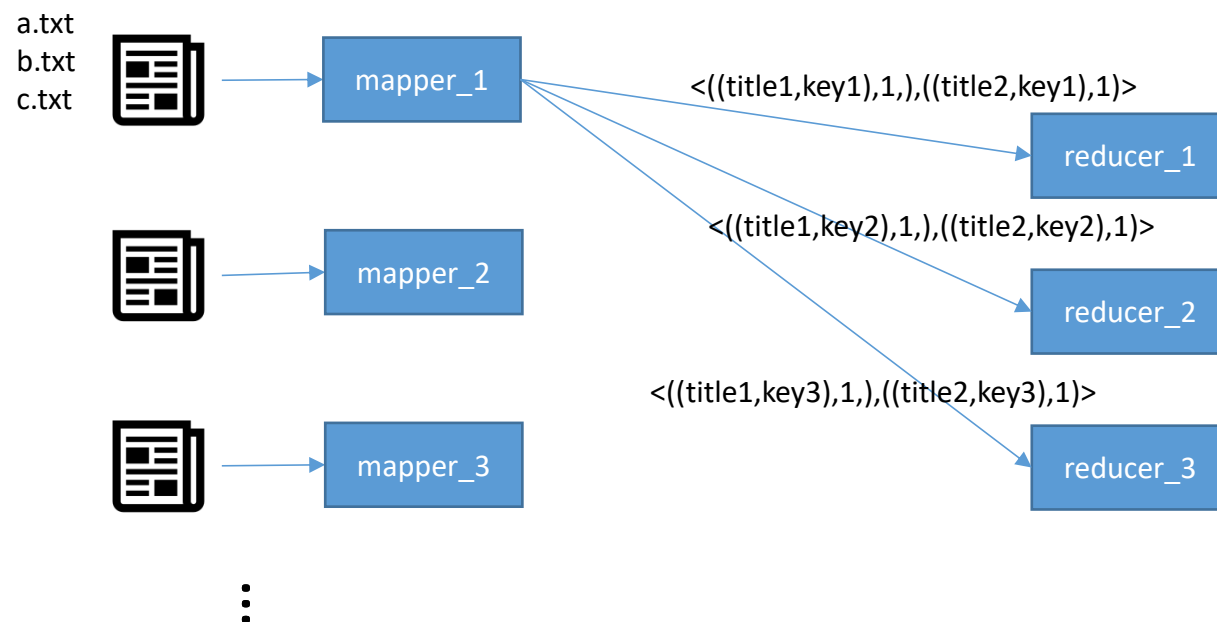
◆ MapReduce参考算法流程



关于mapper结果的分配: 最简单的版本可以是每3个mapper输出的文件作为一个reducer的输入文件。

但是, 当每个mapper输入的文件数量差距很大时, 不同reducer的工作量差异可能会很大。

◆ 进阶：使用shuffle过程。

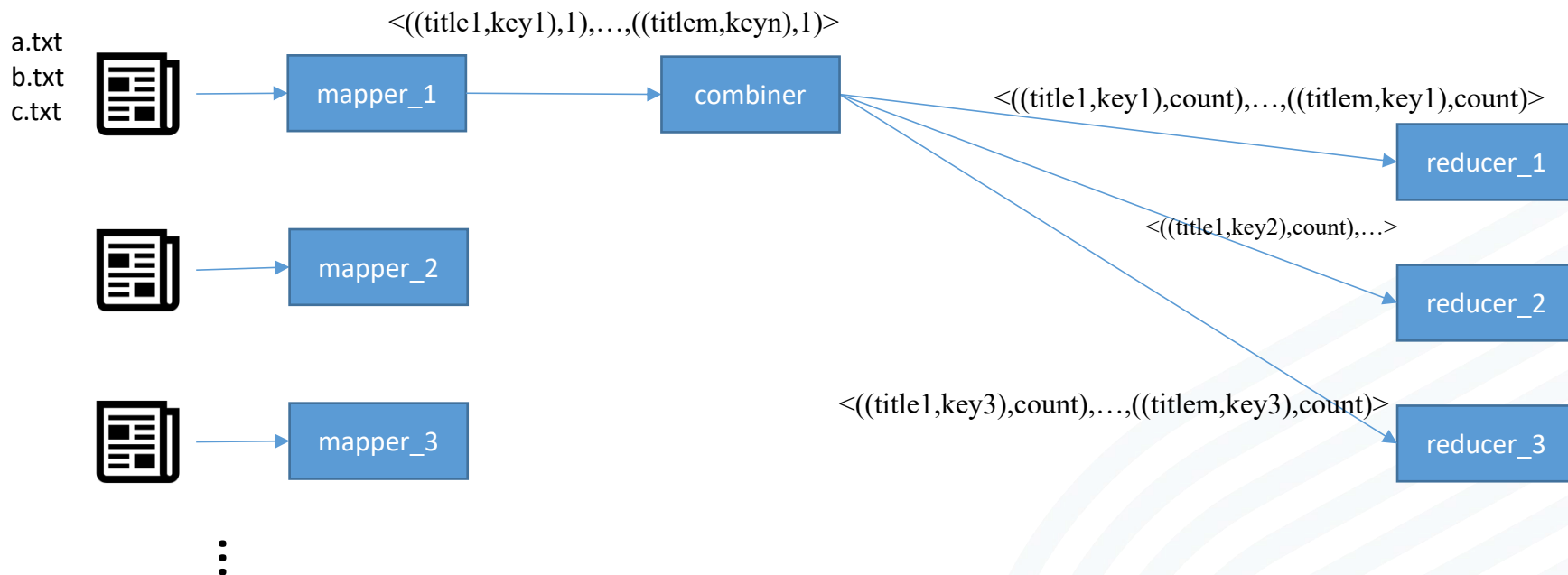


使用shuffle，一个mapper将输出平分为多份，分给多个reducer，这样每个reducer的工作量大致相同。

可以通过hash来将keyword分配到reducer上。

◆ 进阶：使用combine过程。

- 注意到在上述方法中，mapper到reducer的传输开销较大，key_list中可能包含很多重复的键字，每个mapper可以通过combiner来压缩传输开销：





华中科技大学
计算机科学与技术学院
School of Computer Science & Technology, HUST

四 验收流程

四 验收流程



- 统计结果是否正确;
- 验收时对代码的大致解释;
- 验收时的提问与回答。