

暑期实训学习报告

1813012 柳鑫

一. 学习内容

- ① MapReduce 工作原理
- ② MapReduce 的类型与格式

二. 学习结果

- ① MapReduce 工作原理

(1) 分而治之

对相互之间不具有计算依赖关系的大数据,实现并行最自然的办法就是采取分而治之的策略。

(2) Mapper 和 Reduce

MPI 等并行计算方法缺少高层并行编程模型,程序员需要自行指定存储、计算、分发等任务,为了克服这一缺陷,MapReduce 借鉴了 Lisp 函数式语言中的思想,用 Map 和 Reduce 两个函数提供了高层的并发编程模型抽象。

(3) 统一架构

MPI 等并行计算方法缺少统一的计算框架支持,程序员需要考虑数据存储、划分、分发、结果收集、错误恢复等诸多细节。为此,MapReduce 设计并提供了同意的计算框架,为程序员隐藏了绝大多数系统层面的处理系统。

② MapReduce 的类型与格式

(1) 输入格式

1. 输入分片
2. InputFormat
3. FileInputFormat 类
 - A. CombineFileInputFormat
 - B. TextInputFormat
 - C. KeyValueTextInputFormat
 - D. NLineInputFormat
 - E. SequenceFileInputFormat
 - F. 多种输入

(2) 输出格式

1. 文本输出
2. 二进制输出
3. 多个输出
4. 延时输出