Google MapReduce到底解决什么问题?

阅读 0 收藏 0 2018-12-03
原文链接：click.aliyun.com

> 搞架构的人，Google的架构论文是必看的，但好像大家都不愿意去啃英文论文。故把自己的读书笔记，加入自己的思考，分享给大家。

第二篇，Google MapReduce架构启示（上）。

很多时候，定义清楚问题比解决问题更难。

什么是MapReduce?

它不是一个产品，而是一种解决问题的思路，它有多个工程实现，Google在论文中也给出了它自己的工程架构实现。

MapReduce这个编程模型解决什么问题?

能够用分治法解决的问题，例如：

- 网页抓取
- 日志处理
- 索引倒排
- 查询请求汇总
- ...

画外音：能够发现，现实中有许多基于分治的应用需求。

为什么是Google，发明了这个模型?

Google网页抓取，分析，倒排的多个应用场景，当时的技术体系，解决不了Google大数据量高并发量的需求，Google被迫进行技术创新，思考出了这个模型。

画外音：谁痛谁想办法。

为什么MapReduce对“能够用分治法解决的问题”特别有效?

画外音：分治法详见《分治法与减治法》。

Google MapReduce为什么能够成功?

Google为了方便用户使用系统，提供给了用户很少的接口，去解决复杂的问题。
(1) Map函数接口：处理一个基于key/value(后简称kv)的成对(pair)数据集合，同时也输出基于kv的数据集合；
(2) Reduce函数接口：用来合并Map输出的kv数据集合；

画外音：MapReduce系统架构，能在大规模普通PC集群上实现并行处理，和GFS等典型的互联网架构类似。

用户仅仅关注少量接口，不用关心并行、容错、数据分布、负载均衡等细节，又能够解决很多实际的问题，还有这等好事！

能不能举一个例子，说明下MapReduce的Map函数与Reduce函数是如何解决实际问题的?

举例：假设要统计大量文档中单词出现的个数。

Map

输入KV：pair(文档名称，文档内容)

输出KV：pair(单词，1)

画外音：一个单词出现一次，就输出一个1。

Reduce

输入KV：pair(单词，1)

输入KV：pair(单词，总计数)

以下是一段伪代码，
Map (list<pair($doc_name, $doc_content)>){

foreach(pair in list)

```
}
```

画外音：如果有多个Map进程，输入可以是一个pair，不是一个list。

```
Reduce(list<pair($word, $count)>){// 大量(单词,1)
    map<string,int> result;
    foreach(pair in list)
        result[$word] += $count;

    foreach($keyin result)
        echo pair($key, result[$key]); // 输出list<k,v>
}
```

画外音：即使有多个Reduce进程，输入也是list<pair>，因为它的输入是Map的输出。

最早在单机的体系下计算，输入数据量巨大的时候，处理很慢。如何能够在短时间内完成处理，很容易想到的思路是，将这些计算分布在成百上千的主机上， 但此时，会遇到各种复杂的问题，例如：

- 并行计算
- 数据分发
- 错误处理
- 集群通讯
- ...

这些综合到一起，就成为了一个困难的问题，这也是Google MapReduce工程架构要解决的问题，也就是下一章将要分享的问题，敬请期待。

思路比结论更重要。

原文发布时间为：2018-11-29

Google 架构 负载均衡 产品

相关热门文章

【码农打怪升级之路】行走江湖，你需要解锁哪些技能包？【石杉的架构笔记】
石杉的架构笔记 8 2

消息中间件 RocketMQ 源码解析 —— 调试环境搭建
芋道源码_以德服人_不服就干 3

互联网大厂Java面试题：使用无界队列的线程池会导致内存飙升吗？【石杉的架构笔记】
石杉的架构笔记 63 8

日活亿级的负载均衡架构如何搭建?
老錢 30 3

一年后又来了， Cloudopt AdBlocker - 集拦截广告、安全、网页加速一身的神器
t-baby 5 2

评论

输入评论...