

03 | 大规模数据处理初体验：怎样实现大型电商热销榜？

2019-04-22 蔡元楠 来自北京

《大规模数据处理实战》



你好，我是蔡元楠。

今天我要与你分享的主题是“怎样实现大型电商热销榜”。

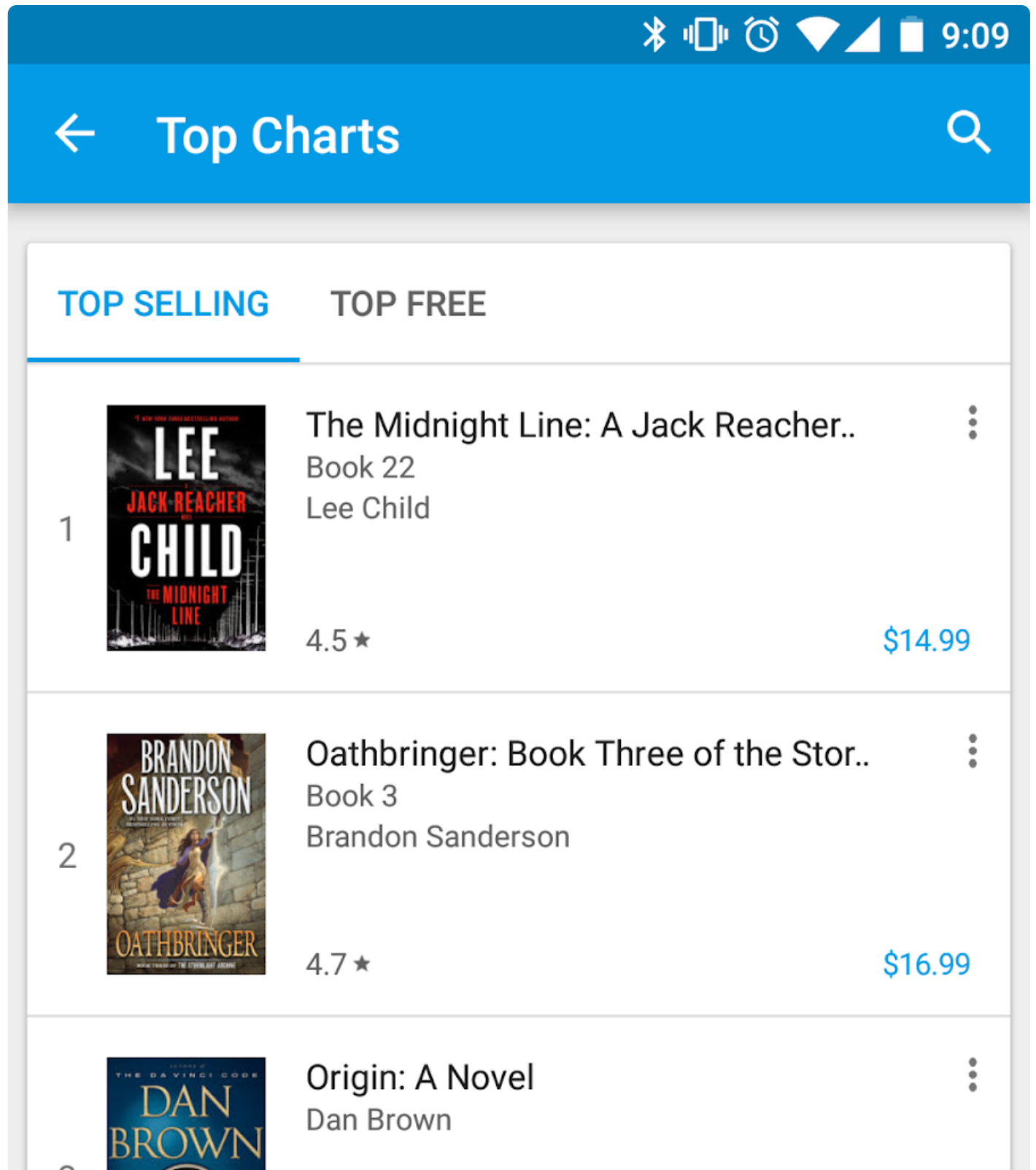
我在 Google 面试过很多优秀的候选人，应对普通的编程问题 coding 能力很强，算法数据结构也应用得不错。

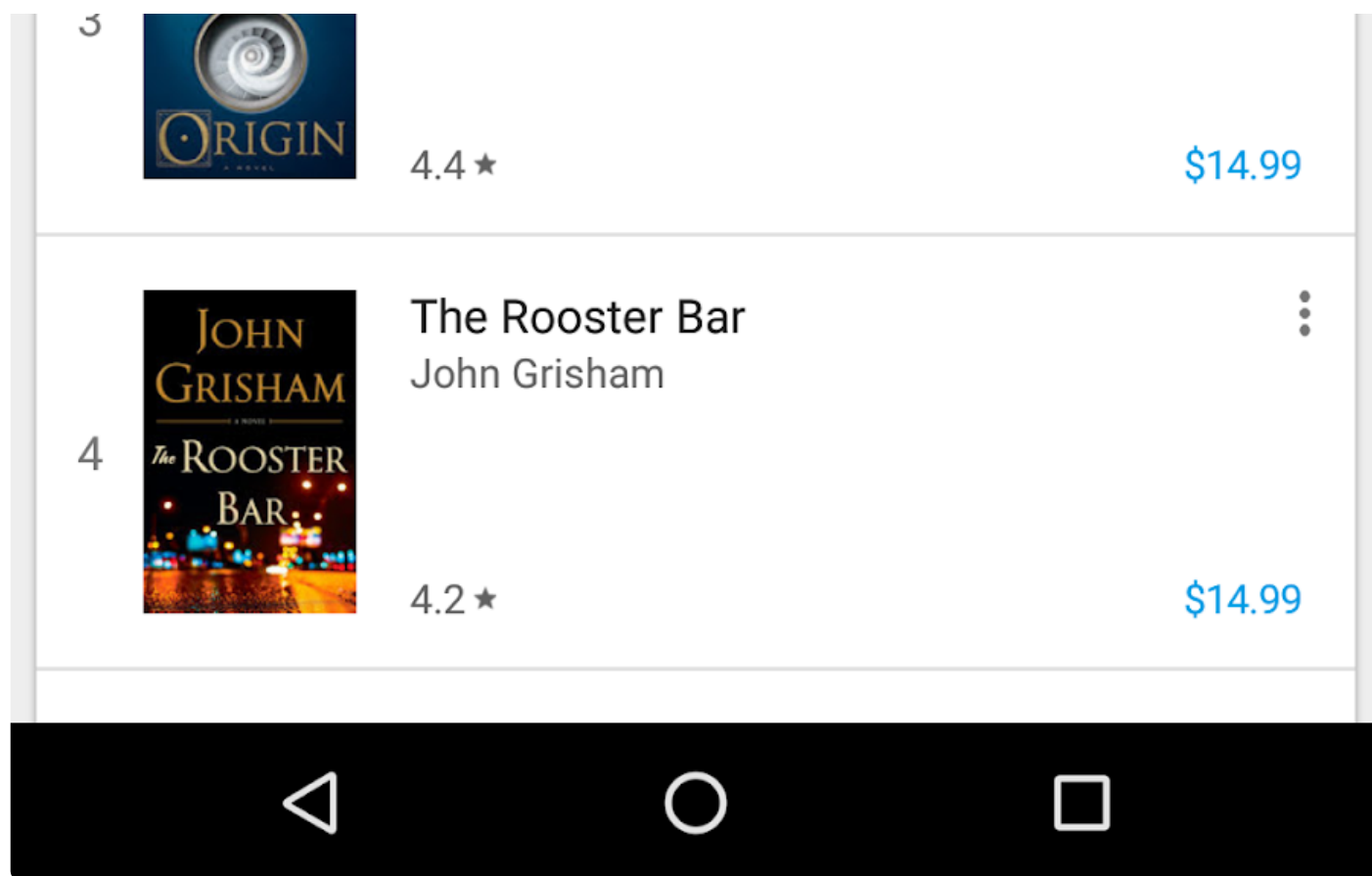
可是当我追问数据规模变大时该怎么设计系统，他们却说不出所以然来。这说明他们缺乏必备的规模增长的技术思维（mindset of scaling）。这会限制这些候选人的职业成长。

因为产品从 1 万用户到 1 亿用户，技术团队从 10 个人到 1000 个人，你的技术规模和数据规模都会完全不一样。

今天我们就以大型电商热销榜为例，来谈一谈从 1 万用户到 1 亿用户，从 GB 数据到 PB 数据系统，技术思维需要怎样的转型升级？

同样的问题举一反三，可以应用在淘宝热卖，App 排行榜，抖音热门，甚至是胡润百富榜，因为实际上他们背后都应用了相似的大规模数据处理技术。





真正的排序系统非常复杂，仅仅是用来排序的特征（features）就需要多年的迭代设计。

为了便于这一讲的讨论，我们来构想一个简化的玩具问题，来帮助你理解。

假设你的电商网站销售 10 亿件商品，已经跟踪了网站的销售记录：商品 id 和购买时间 {product_id, timestamp}，整个交易记录是 1000 亿行数据，TB 级。作为技术负责人，你会怎样设计一个系统，根据销售记录统计去年销量前 10 的商品呢？

举个例子，假设我们的数据是：

product_id	timestamp
1	1553721167
2	1553721199
3	1553721220
1	1553721241

我们可以把热销榜按 product_id 排名为：1, 2, 3。

小规模的经典算法

如果上过极客时间的《数据结构与算法之美》，你可能一眼就看出来，这个问题的解法分为两步：




第一步，统计每个商品的销量。你可以用哈希表（hashtable）数据结构来解决，是一个 $O(n)$ 的算法，这里 n 是 1000 亿。

第二步，找出销量前十，可以用经典的 Top K 算法，也是 $O(n)$ 的算法。

如果你考虑到了这些，先恭喜你答对了。

在小规模系统中，我们确实完全可以用经典的算法简洁漂亮地解决。以 Python 编程的话可能是类似这样的：

 复制代码

```
1 def CountSales(sale_records):
2     """Calculate number of sales for each product id.
3
4     Args:
5         sales_records: list of SaleRecord, SaleRecord is a named tuple,
6             e.g. {product_id: "1", timestamp: 1553721167}.
7     Returns:
8         dict of {product_id: num_of_sales}. E.g. {"1": 1, "2": 1}
9     """
10    sales_count = {}
11    for record in sale_records:
12        sales_count[record[product_id]] += 1
13
14    return sales_count
15
16 def TopSellingItems(sale_records, k=10):
17     """Calculate the best selling k products.
18
19     Args:
20         sales_records: list of SaleRecord, SaleRecord is a named tuple,
21             e.g. {product_id: "1", timestamp: 1553721167}.
22         K: num of top products you want to output.
23     Returns:
24         List of k product_id, sorted by num of sales.
25     """
26    sales_count = CountSales(sale_records)
27    return heapq.nlargest(k, sales_count, key=sales_count.get)
```

但在一切系统中，随着尺度的变大，很多方法就不再适用。

比如，在小尺度经典物理学中适用的牛顿力学公式是这样的：

$$\mathbf{F} = m\mathbf{a}$$

这在高速强力的物理系统中就不再适用，在狭义相对论中有另外的表达。

$$\mathbf{F} = \frac{\gamma^3 m_0 (\mathbf{v} \cdot \mathbf{a})}{c^2} \mathbf{v} + \gamma m_0 \mathbf{a}$$

在社会系统中也是一样，管理 10 人团队，和治理 14 亿人口的国家，复杂度也不可同日而语。

具体在我们这个问题中，同样的 Top K 算法当数据规模变大会遇到哪些问题呢？

第一，内存占用。

对于 TB 级的交易记录数据，很难找到单台计算机容纳那么大的哈希表了。你可能想到，那我不要用哈希表去统计商品销售量了，我把销量计数放在磁盘里完成好了。

比如，就用一个 1000 亿行的文件或者表，然后再把销量统计结果一行一行读进后面的堆树 / 优先级队列。理论上听起来不错，实际上是否真的可行呢，那我们看下一点。

第二，磁盘 I/O 等延时问题。

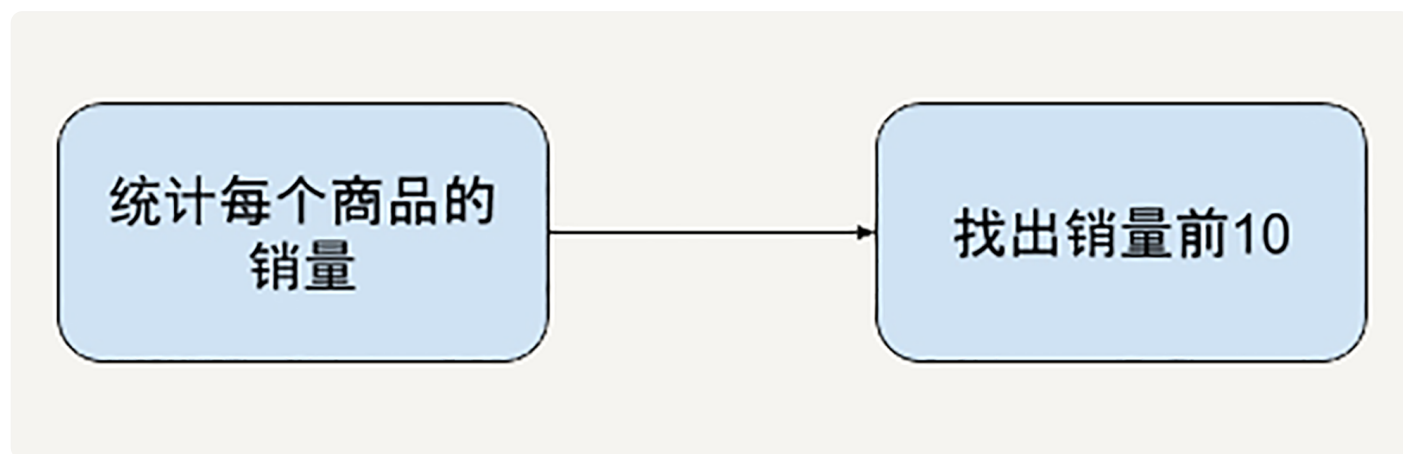
当数据规模变大，我们难以避免地需要把一些中间结果存进磁盘，以应对单步任务出错等问题。一次磁盘读取大概需要 10ms 的时间。

如果按照上一点提到的文件替代方法，因为我们是一个 $O(n * \log k)$ 的算法，就需要 $10\text{ms} * 10^9 = 10^7 \text{ s} = 115$ 天的时间。你可能需要贾跃亭附体，才能忽悠老板接受这样的设计方案了。

这些问题怎么解决呢？你可能已经想到，当单台机器已经无法适应我们数据或者问题的规模，我们需要横向扩展。

大规模分布式解决方案

之前的思路依然没错。但是，我们需要把每一步从简单的函数算法，升级为计算集群的分布式算法。

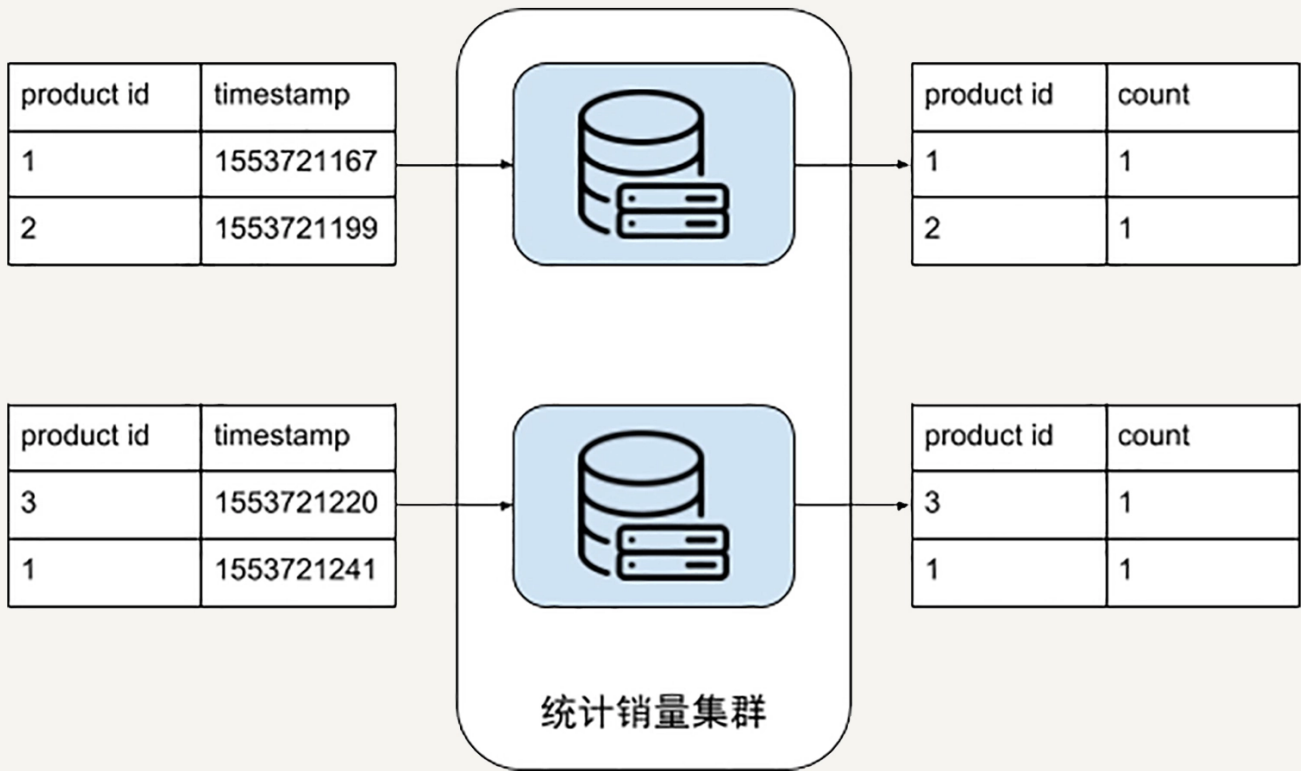


统计每个商品的销量

我们需要的第一个计算集群，就是统计商品销量的集群。

例如，1000 台机器，每台机器一次可以处理 1 万条销售记录。对于每台机器而言，它的单次处理又回归到了我们熟悉的传统算法，数据规模大大缩小。

下图就是一个例子，图中每台机器输入是 2 条销售记录，输出是对于他们的本地输入而言的产品销量计数。



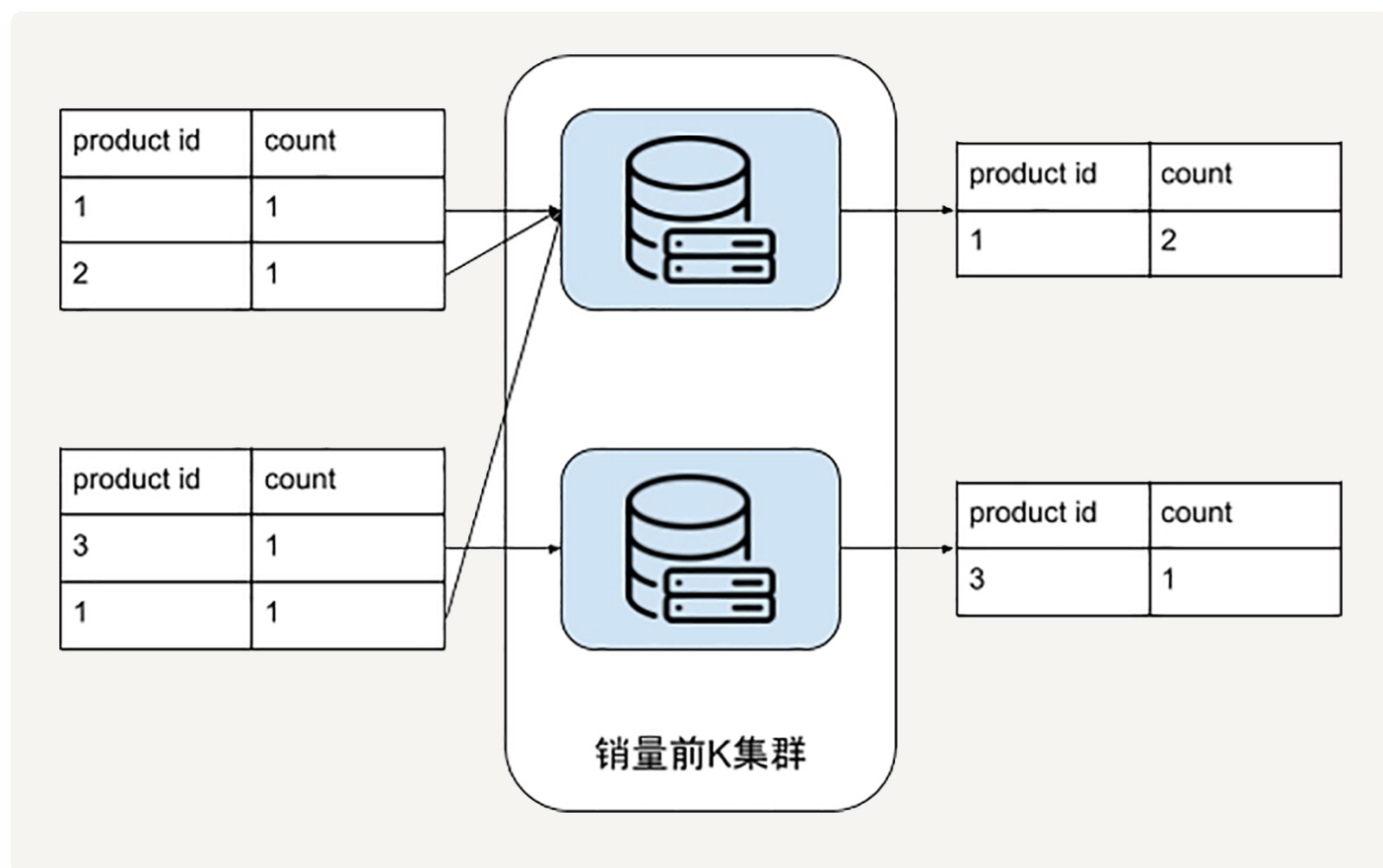
找出销量前 K

我们需要的第二个计算集群，则是找出销量前十的集群。

这里我们不妨把问题抽象一下，抽象出是销量前 K 的产品。因为你的老板随时可能把产品需求改成前 20 销量，而不是前 10 了。

在上一个统计销量集群得到的数据输出，将会是我们这个处理流程的输入。所以这里需要把分布在各个机器分散的产品销量汇总出来。例如，把所有 `product_id = 1` 的销量全部叠加。

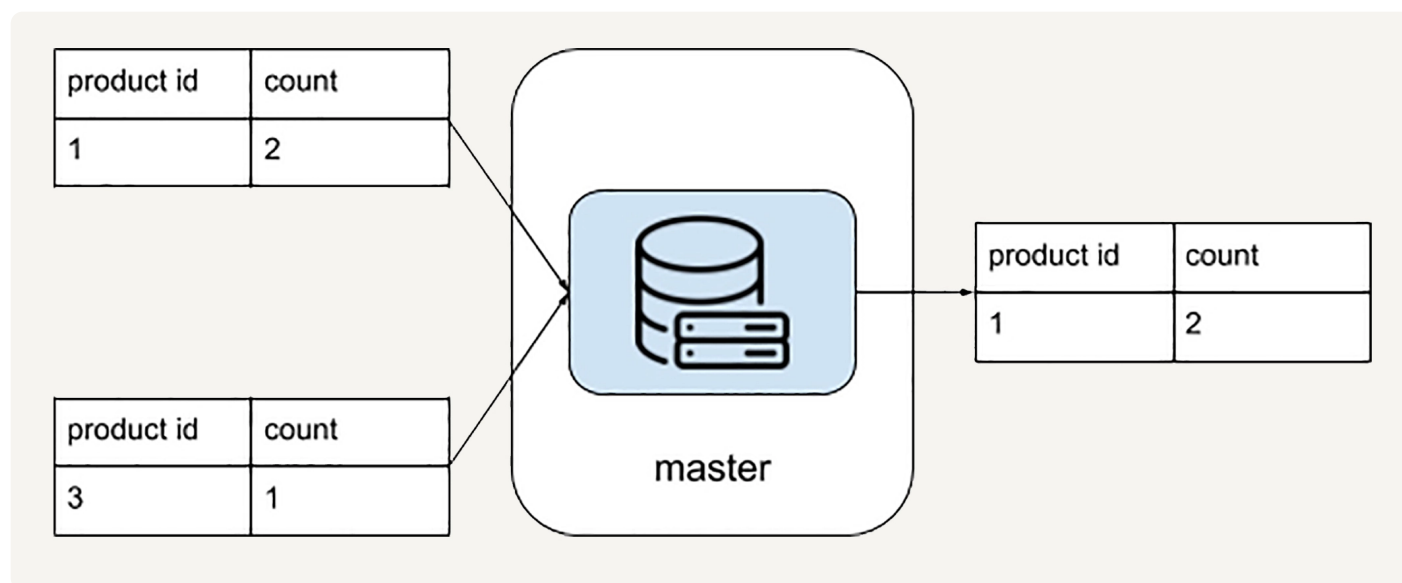
下图示例是 $K = 1$ 的情况，每台机器先把所有 `product_id = 1` 的销量叠加在了一起，再找出自己机器上销量前 $K = 1$ 的商品。可以看到对于每台机器而言，他们的输出就是最终排名前 $K = 1$ 的商品候选者。



汇总最终结果

到了最后一步，你需要把在“销量前 K 集群”中的结果汇总出来。也就是说，从所有排名前 $K=1$ 的商品候选者中找出真正的销量前 $K=1$ 的商品。

这时候完全可以用单一机器解决了。因为实际上你汇总的就是这 1000 台机器的结果，规模足够小。



看到这里，你已经体会到处理超大规模数据的系统是很复杂的。

当你辛辛苦苦设计了应对 1 亿用户的数据处理系统时，可能你就要面临另一个维度的规模化（scaling）。那就是应用场景数量从 1 个变成 1000 个。每一次都为不同的应用场景单独设计分布式集群，招募新的工程师维护变得不再“可持续发展”。

这时，你需要一个数据处理的**框架**。

大规模数据处理框架的功能要求

在第二讲“MapReduce 后谁主沉浮：怎样设计现代大规模数据处理技术”中，我们对于数据处理**框架**已经有了基本的方案。

今天这个实际的例子其实为我们的设计增加了新的挑战。

很多人面对问题，第一个想法是找有没有开源技术可以用一下。

但我经常说服别人不要先去看什么开源技术可以用，而是从自己面对的问题出发独立思考，忘掉 MapReduce，忘掉 Apache Spark，忘掉 Apache Beam。

如果这个世界一无所有，你会设计怎样的大规模数据处理框架？你要经常做一些思维实验，试试带领一下技术的发展，而不是永远跟随别人的技术方向。


在我看来，两个最基本的需求是：

1. 高度抽象的数据处理流程描述语言。作为小白用户，我肯定再也不想一一配置分布式系统的每台机器了。作为框架使用者，我希望框架是非常简单的，能够用几行代码把业务逻辑描述清楚。
2. 根据描述的数据处理流程，自动化的任务分配优化。这个框架背后的引擎需要足够智能，简单地说，要把那些本来手动配置的系统，进行自动任务分配。

那么理想状况是什么？对于上面的应用场景，我作为用户只想写两行代码。

第一行代码：


```
1 sales_count = sale_records.Count()
```

 复制代码

这样简单的描述，在我们框架设计层面，就要能自动构建成上文描述的“销量统计计算集群”。

第二行代码

```
1 top_k_sales = sales_count.TopK(k)
```

 复制代码

这行代码需要自动构建成上文描述的“找出销量前 K 集群”。

看到这里，你能发现这并不复杂。我们到这里就已经基本上把现代大规模数据处理架构的顶层构造掌握了。而背后的具体实现，我会在后面的专栏章节中为你一一揭晓。

小结

这一讲中，我们粗浅地分析了一个电商排行榜的数据处理例子。

从 GB 数据到 TB 数据，我们从小规模算法升级到了分布式处理的设计方案；从单一 TB 数据场景到 1000 个应用场景，我们探索了大规模数据处理框架的设计。

这些都是为了帮助你更好地理解后面所要讲的所有知识。比如，为什么传统算法不再奏效？为什么要去借助抽象的数据处理描述语言？希望在后面的学习过程中，你能一直带着这些问题出发。

思考题

在你的工作中，有没有随着数据规模变大，系统出问题的情况，你又是怎么解决的？

欢迎你把自己的想法写在留言区，与我和其他同学一起讨论。

如果你觉得有所收获，也欢迎把文章分享给你的朋友。

© 版权归极客邦科技所有，未经许可不得传播售卖。页面已增加防盗追踪，如有侵权极客邦将依法追究其法律责任。

精选留言 (50)



青石

2019-04-22

好多年前还未接触大数据时，写过日志采集统计各接口请求报表及puv的脚本，经历了几个阶段。

1. 最初是汇总所有日志到一台服务器，在处理日志，测试环境没问题，上生产跑起来就几个小时。
2. 后来分到Web服务器各自处理数据，时间缩短了，但是汇总数据偶尔会有问题。
3. 将数据写入到数据库，解决汇总数据问题。但是单表数据量过大，统计又很慢。
4. 按天分表解决数据量问题，最后就这么一直运行下去了。

这段经历其实很普通，但也确实让我更轻松的学习和理解大数据。当我学到mapreduce内容的时候，回忆起这段经历，让我很容易就接受了mapreduce的分治思想。

就像看到hbase的时候，我的理解它就是在实现数据的寻址、不断的拆分/合并表，但是原来的人工操作变成现在自动化操作。

共 1 条评论 >

👍 43



Liu C.

2019-04-22

有一次处理一个非常高维的feature矩阵，要给它降维，但手头的电脑cpu和内存都不够好。于是我用了非常hack的手段：先使用random projection算法降低一定维度，这是一个纯矩阵乘法，可以分块放入内存计算。之后剩余的维度还是有些大，于是我把feature拆成几组，对每组分别做pca，之后再选出每组最大的主成分拼起来，就完成了降维。

作者回复：谢谢你的经验分享！



👍 34



bwv825

2019-04-27

Top 1 的情况，只统计每台机器的top 1是不是可能会不准确呢？比如数据按时间段分片，某个商品销量很大很稳定，累计总数第一但很少是top 1, 因为各个时间段都有不同的爆款...

共 9 条评论 >

👍 17



Mr Zhuo

2019-04-22

老师好，我目前是做NLP落地的，本来是作为补充知识学的这个专栏，但是学了这几节后发现这个方向很有潜力，也很感兴趣。另外由于你们google的BERT横空出世，感觉NLP方向的个人发展有些迷茫，所以想请问老师，对于专栏内容和NLP的结合，在未来发展有没有好的建议呢？



👍 14



孙稚昊

2019-04-23

数据量一大，最常见的问题除了各种exception，就是key 值分布不均衡。电商一般都是长尾的，少量的item 占据大多数购买量，很容易发送数据倾斜，需要设计更新的hash-sharding 方法



👍 11



Kev1n

2019-04-22

个人经验，拆分，复制，异步，并行，是大规模数据处理和应用架构的常见手段，一致性根据

业务场景适当妥协



11



hua168

2019-04-22

分解法...像剁鱼那样，一条一口吃不下就切成块，块一口吃还大，有风险，再就再用筷子分小...

关键问题是怎么切，切多大？怎么不全切碎，让它完整的，让人知道是条鱼 😊

作者回复：你这比喻很屌



10



乘坐Tornado的线程魔...

2019-04-22

作者好！找出前K个集群小节里面的第一个计算集群的第二个节点（机器），是否应该像第一个节点一样计算product_id=1的所有记录。文中图示貌似只有第一个节点计算了。请作者查证。

作者回复：这个图里面是按照product_id分组了，所以所有product id =1的都归第一个机器



10



孙稚昊

2019-04-23

我们在做商品订单统计的时候，会按itemid + order year + order month 对订单做hash来做group 的 key，分割成更小块，防止popular item 堆积造成的瓶颈

共 2 条评论 >

9



leeon

2019-04-24

大规模的topk在计算过程中很容易引发数据倾斜的问题，在实际业务里，计算的优化是一方面，有时候从数据层面去优化也会有更好的效果，以榜单为例，可以在时间维度和地域为度去拆解数据，先小聚再大聚



7



2019-04-23

最初，GPS数据以文件形式存储在盘阵中，数据增长达到TB级别后，考虑到性能和成本以及可扩展性，系统迁移到HDFS中，离线任务用MR，在线查询采用HBSE，现在，数据PB级别后，发现热点数据hbase成本太高，系统迁移到时序数据库，专供线上实时查询，同时，实时分析采用storm，批处理用spark。其实，很多情况下，采用什么技术，成本具有决定性因素



👍 6



zhihai.tu

2019-04-22

有一个项目，试点的时候由于用户访问量小，传统负载均衡F5下连6台应用服务器访问为啥问题。后续推广后，由于访问量出现了50倍以上的增加，前台响应慢，服务器也出现内存溢出等问题。后续采用了docker容器技术，从应用服务器上抽取出并发访问较高的服务模块，单独部署服务层，支持横向扩展以及在线扩容，较好的解决了问题。



👍 5



涵

2019-04-22

做传统数仓时，使用oracle数据库，随着数据量增大会需要使用到分区。分区需要思考使用哪个属性来分，分成多大的区间合适。另外，当视图很大时，有时查询很慢，会使用物化视图的方法。

作者回复: 谢谢你的经验分享!



👍 4



JohnT3e

2019-04-22

数据倾斜，导致任务运行时间超出预期，这个时候就需要对数据做一些分析和采样，优化shuffle。任务出错后，调试周期变长，这个目前没有很好的解决。不过，之前看flumejava论文，其采用了缓存不变结果来加快调试周期。另外，就是集群规模增大，后期运维的问题了

作者回复: 谢谢你的经验分享!



👍 5



Daryl

2019-04-29

作者其实关于top k没描述清楚，虽然我明白他的意思，因为我了解这边，但是对于没有了解的同学会有点晕乎



👍 3



乘坐Tornado的线程魔...

2019-04-23

顺便复习了王争老师的《数据结构与算法》，看到Top算法的时间复杂度准确来讲应该是 $O(n\log K)$

作者回复: 这里K远小于n，写成 $O(n)$ 没有问题



👍 3



Charles.Gast

2019-04-22

数据不数据什么的无所谓，我就想听听那个力学公式的讲解㊟

作者回复: 哈哈



👍 3



哈哈

2019-05-10

将大规模数据拆解到多台机器处理，还应该用一定的规则哈希到每台机器吧



👍 2



朱同学

2019-04-26

实际上传统服务也是这样，业务初期我们一台物理机，后面又是三台物理机，做的反向代理小集群，到现在几个机柜做了虚拟化，数据库也做了读写分离，说到底就是集群化处理

作者回复: 谢谢你的经验分享!



👍 2



hufox

2019-04-24

以前做订单系统的时候，由于数据量没有那么大，没有考虑到大规模数据处理问题，但是一旦数据量上来了，统计查询都很慢，今天阅读了老师这一讲，原来可以这样设计处理大规模数据问题，涨姿势了！继续学习！

作者回复: 谢谢支持！



2