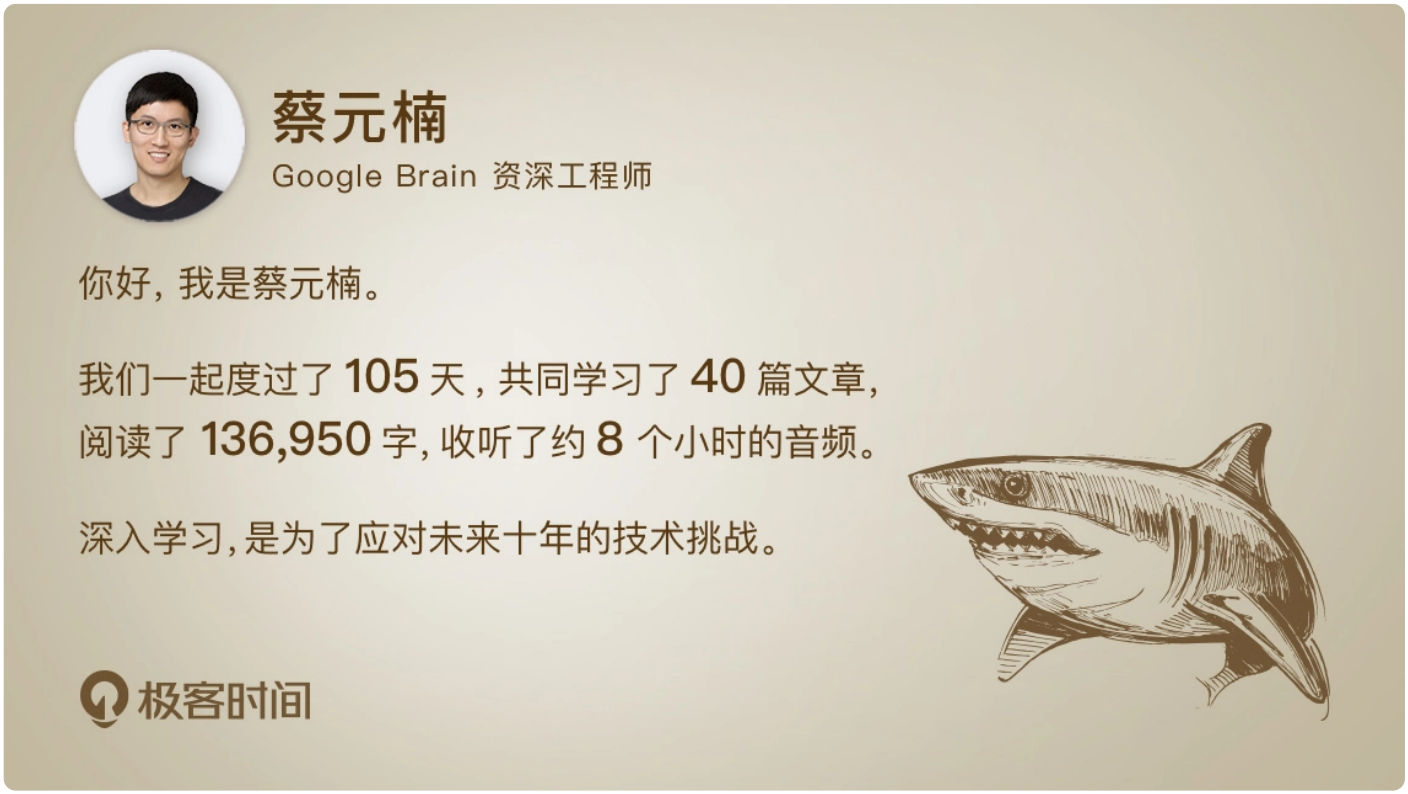
# 结束语 | 世间所有的相遇，都是久别重逢

2019-07-29 蔡元楠 来自北京

《大规模数据处理实战》




今天和往常一样。我开车沿着硅谷的 101 公路下班，101 还是一样的堵。

今天和往常又不一样。比如，今天午饭的话题不一样。往日我会利用午饭或者喝咖啡的时间和同事一起讨论一些极客时间专栏读者提出的问题。而且，今天回到家里也不需要赶着写稿。如果是往日，编辑催稿的微信早就已经堆积成了 99+ 未读消息。

## 专栏成功的指标 = 所有读者收获之和

我还清楚地记得，4 月 15 日（北京时间），因为时差原因，我在凌晨 2、3 点等待着专栏上线，守着手机屏幕，看新的读者加入。我在极客时间的留言管理页面一遍一遍下拉刷新，期待着去回复包括你在内的读者留言。说不定我一个人就给极客时间的服务器 QPS 增加了 1 或者 2（笑）。

写作的过程中，我也碰到过很多困难。

专栏占用了很多时间，有时候为了查证一个技术点我会研读 Apache Beam 代码到深夜，女朋友甚至因此和我吵过架。不仅是写作的时间需要去平衡，我的全职工作和生活也需要。在专栏更新过半的时候，因为家人身体原因，我必须一直往返于湾区和德州，没有完整的时间录音，不得不更换了主播来录音。

为此，我也在这里，再次和你道歉。

当然，也有写作时间之外的阻力。我在开篇词中对专栏的内容要求是，每一篇专栏都是原汁原味的硅谷技术分享。这也带来了一个意外的问题。

第 1 讲“为什么 MapReduce 会被硅谷一线大厂淘汰”一经在知乎分享，就一跃成为 MapReduce 话题点赞排名第一的文章。它的影响力超过了 Google 内容发行部门的预计，以至于负责 Google 内容审查的部门找到我，要求一起审查专栏中还未发布的内容。

这样的小插曲其实我也没有预料到。在经历了 14 天详细的审查之后，我的专栏终于通过审核，被认定为是开源技术分享，而并不涉及商业机密。

专栏写作期间，有时候和朋友吃个饭我都不得不提前告辞：“今晚需要回家写稿”。朋友都问我：“为什么你要花这么多时间去写专栏？”

是啊，为什么？

因为你。专栏上线的十几天的时候，在 LinkedIn、微信等各种社交渠道里，有来自 Apple、Uber、思科等各种公司的读者联系到我，和我说：“蔡老师，谢谢你，我看了你的专栏，真的收获很大。”

你的收获，你的成长，才是我如此认真写作的动力。即使订阅量只有 1，我也会为了这 1 份订阅背后的信任，100% 交付每一篇内容。

## 万物皆数据

记得在刚入职 Google 的时候，组里希望我能在了解大组业务的同时，熟悉 Google 内部对于大规模数据处理的整个流程和做法。而组里布置给我的第一个启动项目就是修改一个大型

MapReduce 任务中的一小部分逻辑。

当时我着了迷似地研究组内数据处理的整套流程。

我逐渐发现，Google 内部所使用的 API 和我在上学时自学的 Apache Spark 非常不同。即便需要新加或者修改的逻辑非常多，再加上数据异常的检测（Monitoring）和执行流程日志（Logging）的输出，整个代码加起来也很难超过 100 行。

而我组里的技术领导（Tech Lead）更是告诉我，在查看自己所修改的 Pipeline 的时候，可以利用 DAG 来查看一整套执行流程图，我更是觉得十分神奇。

后来我才慢慢知道，其实在我加入 Google 的时候，Google 已经完全淘汰了以前的 MapReduce Framework，而将 Google Dataflow 的整套思想都运用到了这个内部称为 Flume 的大规模数据处理 Framework 中去。在新入职员工的内部培训中，讲师就说过——Flume 是每个 Noogler 的必修课。

在经历了几年的技术历练之后，我开始成为 Google 的面试官之一。

一般来说，面试题目除了基本的数据结构和算法外，还会涉及到系统设计。其实这些题目或多或少都在考察着我们如何处理转换数据，如何保存数据。在和 Google 不同的组之间进行交流后，我也更加清楚地了解到，其实每个组都在做着数据处理的任务。

我们自己的生活难道不也是完完全全被大规模数据处理充斥着的吗？

例子太多了，日常生活购物网站（淘宝、Amazon）在处理着所有客户的交易商品数据，平时使用的支付软件（支付宝、Venmo）在处理着我们的支付数据，社交软件（微博、Facebook 和 Twitter）在处理着大规模社交信息流数据。

其实我们一直都生活在一个被大规模数据处理所包围着环境中，我相信等到 5G 全面到来，这种现象会更加明显。而我也越来越觉得，学习好大规模数据处理真的是一门必修课。

## 我的下一步

虽然我们对于大规模数据处理的学习已经告一段落，但是我坚信“万物皆数据”，坚信数据会改变人类生活。

接下来，我会回收更多的时间精力，继续从事人工智能医疗健康的研发工作。除此之外，我还会参与录制 Google 官方的 TensorFlow 视频课程。

感谢你在这三个月里的努力学习，希望你能学有所得，也希望有机会能与你再次重逢！

蔡元楠

Google Brain 资深工程师

不知道在学习过程中，你有哪些体会和评价？
这里有一份专栏调查问卷，邀请你填写。

**在8月5日前提交，**
**极客时间赠送给你专属优惠券。**

我们一起继续成长！

去提交



## 精选留言 (42)

**梅亮宏@创造力**

2019-07-30

老师是有信仰的一个人，任何愿意分享技术都是很有魅力的。你女朋友应该感到自豪：）谢谢三个月的陪伴！Good luck in your new journey! Hopefully see you again  on Geek's Time! Cheers!

13

**微思**

2019-07-29

赞

13

**兆熊**

2020-04-18

我用思维导图的形式对专栏的内容进行了总结，欢迎阅读： https://blog.csdn.net/zhouzxi/article/details/105594026?utm_source=app

5

**cotter**

2019-07-30

首个追完的专栏，感谢蔡老师

3

**never leave**

2019-07-30

感谢老师的辛苦付出

3

**天空只能仰望?**

2019-07-30

老师，你好，请教一下beam运行如何管理应用的中间状态，类似于flink checkpoint ？

作者回复: 谢谢你的提问！像Checkpoint和Drain这种概念Beam现在暂时还不支持，不过我相信在roadmap中。

3

**inzaghi**

2020-02-14

hi，老师，你好。你有没有大数据处理+用户画像+推荐系统方向的教程、或者书籍推荐。谢谢！

2

**¾**

2019-07-30

收获颇多 感谢！

2

**陈凯枫**

2019-07-30

感谢蔡老师的辛勤付出！通过专栏学习，开拓了视野，提高了见识。

2

**Scarsy**

2019-07-30

谢谢老师，老师辛苦了

2

**TKbook**

2019-07-29

终于看完专栏，感谢老师，牺牲这么多时间。。。

2

**三水**

2019-07-29

这是目前唯一追更学习的专栏，谢谢老师！

2

**旭**

2019-07-29

感谢分享

2

**Samlam**

2019-07-29

感谢蔡老师 🙏

2

**kenan**
2019-07-29

老师，诚挚之眼，感人肺腑，我们下一门课程相见。

2



**JensonYao**
2019-07-29

感谢蔡老师!

2

**apollo**
2019-07-29

感谢你!

2

**FengX**
2019-07-30

感谢老师在这一百多天里的辛勤付出!

1

**未知者**
2019-07-29

感谢老师

1

**极客若鸟**
2022-03-11

感谢作者，学完之后，心态发生了很大的变化