“你有多少大数据和Hadoop的经验？”他们问我。我一直在用Hadoop，但很少处理几TB以上的任务。我基本上只是一个大数据新手知道概念，写过代码，但是没有大规模经验。

接下来他们会问：“你能用Hadoop做简单的group by和sum操作吗？”我当然会，但我会说需要看看具体文件格式。

他们给我一个U盘，里面有所有的数据，600MB，对，他们所有的数据。不知道为什么，我用pandas.read\_csv（Pandas是一种Python数据分析库）而不是Hadoop完成了这个任务后，他们显得很不满意。

Hadoop其实是挺局限的。它无非是运行某个通用的计算，用SQL伪代码表示就是： SELECT G(…) FROM table GROUP BY F(…) 你只能改变G和F操作，除非要在中间步骤做性能优化（这可不怎么好玩！）。其他一切都是死的。

（关于MapReduce，之前作者写过一篇“41个词讲清楚MapReduce”，可以参考。）

Hadoop里，所有计算都必须按照一个map、一个group by、一个aggregate或者这种计算序列来写。这和穿上紧身衣一样，多憋得慌啊。许多计算用其他模型其实更适合。忍受紧身衣的唯一原因就是，可以扩展到极大极大的数据集。可你的数据集实际上很可能根本远远够不上那个数量级。

可是呢，因为Hadoop和大数据是热词，世界有一半的人都想穿上紧身衣，即使他们根本不需要。

### ****可我的数据有好几百MB呢！Excel都装不下****

对Excel很大可不是什么大数据。有很多好工具我喜欢用的是基于Numpy的Pandas。它可以将几百MB数据以高效的向量化格式加载到内存，在我已经3年的老笔记本上，一眨眼的功夫，Numpy就能完成1亿次浮点计算。Matlab和R也是很棒的工具。

数百MB数据一般用一个简单的Python脚本逐行读取文件、处理，然后写到了一个文件就行了。

### ****可我的数据有10G呢！****

我刚买了一台笔记本电脑。16G内存花了141.98美元，256GB SSD多收200美元。另外，如果在Pandas里加载一个10GB的csv文件，实际在内存里并没有那么大你可以将 “17284932583” 这样的数值串存为4位或者8位整数，“284572452.2435723”存为8位双精度。

最差情况下，你还可以不同时将所有数据都一次加载到内存里。

### ****可我的数据有100GB/500GB/1TB！****

一个2T的硬盘才94.99美元，4T是169.99。买一块，加到桌面电脑或者服务器上，然后装上PostgreSQL。

### ****Hadoop的适用范围远小于SQL和Python脚本****

从计算的表达能力来说，Hadoop比SQL差多了。Hadoop里能写的计算，在SQL或者简单的Python脚本都可以更轻松地写出来。

SQL是直观的查询语言，没有太多抽象，业务分析师和程序员都很常用。SQL查询往往非常简单，而且一般也很快只要数据库正确地做了索引，要花几秒钟的查询都不太多见。

Hadoop没有任何索引的概念，它只知道全表扫描。而且Hadoop抽象层次太多了我之前的项目尽在应付Java内存错误、内存碎片和集群竞用了，实际的数据分析工作反而没了时间。

如果你的数据结构不是SQL表的形式（比如纯文本、JSON、二进制），一般写一小段Python或者Ruby脚本按行处理更直接。保存在多个文件里，逐个处理即可。SQL不适用的情况下，从编程来说Hadoop也没那么糟糕，但相比Python脚本仍然没有什么优势。

除了难以编程，Hadoop还一般总是比其他技术方案要慢。只要索引用得好，SQL查询非常快。比如要计算join，PostgreSQL只需查看索引（如果有），然后查询所需的每个键。而Hadoop呢，必须做全表扫描，然后重排整个表。排序通过多台机器之间分片可以加速，但也带来了跨多机数据流处理的开销。如果要处理二进制文件，Hadoop必须反复访问namenode。而简单的Python脚本只要反复访问文件系统即可。

### ****可我的数据超过了5TB！****

你的命可真苦只能苦逼地折腾Hadoop了，没有太多其他选择（可能还能用许多硬盘容量的高富帅机器来扛），而且其他选择往往贵得要命（脑海中浮现出IOE等等字样……）。

用Hadoop唯一的好处是扩展。如果你的数据是一个数TB的单表，那么全表扫描是Hadoop的强项。此外的话，请关爱生命，尽量远离Hadoop。它带来的烦恼根本不值，用传统方法既省时又省力。

### ****附注：Hadoop也是不错的工具****

我可不是成心黑Hadoop啊。其实我自己经常用Hadoop来完成其他工具无法轻易完成的任务。（我推荐使用Scalding，而不是Hive或者Pig，因为你可以用Scala语言来写级联Hadoop任务，隐藏了MapReduce底层细节。）我本文要强调的是，用Hadoop之前应该三思而行，别500MB数据这样的蚊子，你也拿Hadoop这样的大炮来轰。