如何在开源项目中学到更多

当前，开源、开放、协作、分享，已成业界共识，开源软件成为软件交付的主要方式。互联网巨头们都在各自领域打造自己的开源生态系统，Google的Android已经是全球最大智能手机生态系统。它的Kubernetes、TensorFlow也有望成为容器时代和机器学习的Android。Facebook、Amazon、Microsoft、国内的互联网巨头也紧随其后，纷纷布局自己的开源生态系统。这股潮流还将继续下去，对于我们大多数人，如何跟上脚步也就变得异常重要，本期话题就试着从以下几个方面与大家进行探讨：

一、懂得协作：维基百科, Apache, GNU/Linux, OpenStack 基金会的成功经验

维基百科大家应该都比较熟悉，经常查资料，它也是我每天用得最多的服务。

维基百科是一个现象，是互联网大规模协作的先驱。

维基百科证实了大规模协作是如何改变一切的，也衍生出一种新的经济形态：维基经济学，它对商业社区产生深远影响。

维基经济学是一门新的科学和艺术：它以四个新法则为基础：开放、对等、共享、全球运作。

最近非常火的共享经济，可以说是维基经济学更大规模的商业实践，GitHub正通过协作方式汇集全球程序员的智慧。

GNU/Linux

自由、开源的代名词，也是有史以来最成功、影响最深远的开源生态系统。

大家平时可关注和使用Linux两大类发行版：Debian/Ubuntu和Redhat/CentOS。

我个人最喜欢的是Debian，因为它的社群契约和愿景，庞大的生态系统(超过51000个软件包)，还因为它的标识。
大家可想想为什么是Debian和炉石传说？

Apache基金会

历史悠久， 因为Apache Web Server而闻名，曾是全球最广泛使用的Web Server。尽管有Nginx和Microsoft的更多选择，但Apache依旧是Web普适性的代表。

https://news.netcraft.com/archives/category/web-server-survey/

Developer	June 2017	Percent	July 2017	Percent	Change
Microsoft	862,255,584	48.80%	940,029,828	53.17%	4.37
Apache	371,461,399	21.02%	315,188,480	17.83%	-3.20
nginx	239,666,345	13.56%	266,041,296	15.05%	1.48
Google	20,136,304	1.14%	20,855,424	1.18%	0.04

ps: Google的服务器数量有多少？这一直是个谜。

Doug Cutting 在 1999编写了 Lucene ，2001加入Apache Jakarta项目，2005成为Apache顶级项目。 Apache Solr 2010加入Lucene子项目，这个一个著名开源搜索引擎项目。2006，Hadoop从Apache Nutch(lucene子项目)启动并开启Hadoop大数据时代。Yahoo在这两个项目扮演重要角色，是最大的贡献者，可惜现在没有Yahoo了。在过去的这10年，开源搜索引擎和大数据在全球范围内大规模部署和运营，并取得巨大成功。

现在围绕Hadoop大数据生态的Apache开源项目有几十个，非常的丰富。几乎可以这样说，你做大数据，一定绕不过Apache项目。

大家可关注两个主要Hadoop发行版，快速体验和了解大数据生态系统。

Hortonworks HDP和Cloudera CDH

OpenStack基金会

2010开源，2017年几乎成为这个星球上发展最快、最大规模的开源项目，仅次于Linux的第二大开源社区。我们都知道云平台在整个互联网和IT工业的重要性，它是基石。OpenStack目前是开源云平台的事实标准，它的发展和想象空间无比巨大。最近与Kubernetes的深度整合，将开启一个全新的时代。

(图片来自：http://www.easystack.cn/escontainer/)

这几个基金会都发展的异常成功，我们需要多想想和仔细思考，如何将自己的职业、工作与开源生态更好的规划下，

下面是一些参考和建议：

《大教堂与集市》是开源运动的《圣经》，颠覆了传统的软件开发思路，影响了整个软件开发领域。
《Apache基金会的运营之道》企业已把自己的开源项目提交Apache并成功孵化升级为顶级项目作为自身开源战略的最佳路线。

一个很核心的原则：ASF不允许企业直接参与Apache项目管理或相关的治理活动。ASF厂商中立，参与仅限于个人，不参杂任何的关系和雇佣状态。

二、学习编程：C++, Java, Python

对语言的选择，我们没有偏见，选择适合自己的就是最好的。

我倾向C++、Java、Python三种语言，其实，我对 Fortran、Lisp、Erlang也独有情钟，对它们充满好奇、敬畏与热爱 :)

先说说为什么选择C++、Java、Python这三种语言，因为它们最具代表性。

ps：我们把C/C++放在一起(常常一同使用)，所以这三大类语言是目前编程语言前三甲，也是Google的三大官方语言。选择它们，有保障。

此外，还有很多语言值得关注，刚才说的C++、Java、Python可以理解为服务端语言。那平时我们用得最多的App，大家可多多关注和实践Swift、Kotlin，它们是iOS和Android的官方语言，代表着未来。同时，它们有趣，也很有价值。

对于Java，我更多想表达的是JVM生态：Clojure、Scala、Kotlin、Java …

关于语言和相关开源项目，可以更多关注GitHub上的

这是优秀、活跃开源项目的大本营：https://github.com/sindresorhus/awesome 非常非常多，够你学习一辈子的。

C/C++多开发系统支撑软件和编程语言，如：数据库：MySQL、MongoDB；编译器：GCC、LLVM；人工智能：TensorFlow、MXNet；编程语言虚拟机：Swift、OpenJDK HotSpot、HHVM

http://wiki.huihoo.com/wiki/C%2B%2B_ecosystem

Java/JVM撑起了大数据整个基础设施，可关注主要发行版：Hortonworks(HDP)、Cloudera(CDP) 它们都融合Spark。

http://wiki.huihoo.com/wiki/Java_ecosystem

Python是数据分析的首选语言，也是系统粘合剂。

Python完成端到端的开发，从云端到万物互联的终端，Python是全栈开发语言。

Python不仅在云平台(OpenStack)、数据分析站稳了脚，也在物联网全栈开发找到了另一片天地，Python也是物联网系统的胶水语言。

参考：

Python 物联网全栈开发经验教训共享

http://wiki.huihoo.com/wiki/Python_ecosystem

此外，因为JVM和Spark，因为静态类型和函数式编程，Scala具有成为数据科学主导语言的潜力。所以，Python之外，Scala你应该更多关注。

IBM在Spark的发展思路是将Spark视为数据分析的操作系统，Spark发行版可多关注：Databricks。
GrowingIO 技术栈是 Scala, Play, Spark, Kafka, HBase, Elasticsearch

这里也产生了一种商业模式：Apache开源项目的分发版，当然这里对团队和开发者有极高的要求。Apache的每一个顶级项目，都可能通过再分发一个商业版本而获得成功。Hortonworks和Cloudera就是这样的成功代表，OpenStack的商业版本也成就了数家公司。

我自己的机器学习和数据挖掘是从Scikit-learn和Weka入手的，我现在虽没有类似GPU和FPGA的设备，做不了Google、Facebook那样的实验，但我可以先从理论和算法入手，Scikit-learn和Weka提供了很便捷的方式。另外Deeplearning4j可多关注，因为JVM是最大的开发者阵营，而DL4J为JVM生态提供了深度学习解决方案，能快速的融合这个生态。而且DL4J的文档也非常棒，可作为深度学习教科书。

最近我们发起了 Deep learning on HDP 开源项目，它是在HDP大数据平台上开发、部署、运营深度学习基础设施，希望更好融合深度学习和大数据。
ps: 项目思路也适合Cloudera(CDP)等其它Hadoop发行版。

我们基于Apache、 OpenStack、 GNU/Linux都有很多的实践。

之前一次电商创业项目，我们基于Apache OFBiz和OpenStack，打造了一个全开放的O2O电商平台，网上商城 + 20间实体店。

这是当时的一份技术分享的资料：百货购OFBiz实践，供大家参考。

最后你可关注我们现在努力推动的三个开源项目(MED)，也欢迎大家参与，所有项目都开源开放：

M3 营销：市场营销和用户增长：增长用户、增加收入。

E3 应用：企业应用和电商平台：理顺企业业务，沉淀业务数据。

D3 数据：数据分析和机器学习：洞悉数据，增强智能。

做互联网和IT行业，我们从事的技术工作，主要要解决的就是算法、数据、算力三个核心问题。

算法，这个比较偏重个人，涉及编程和数学等知识和技能，自己可通过长时间的积累和实践逐渐丰富和提升。

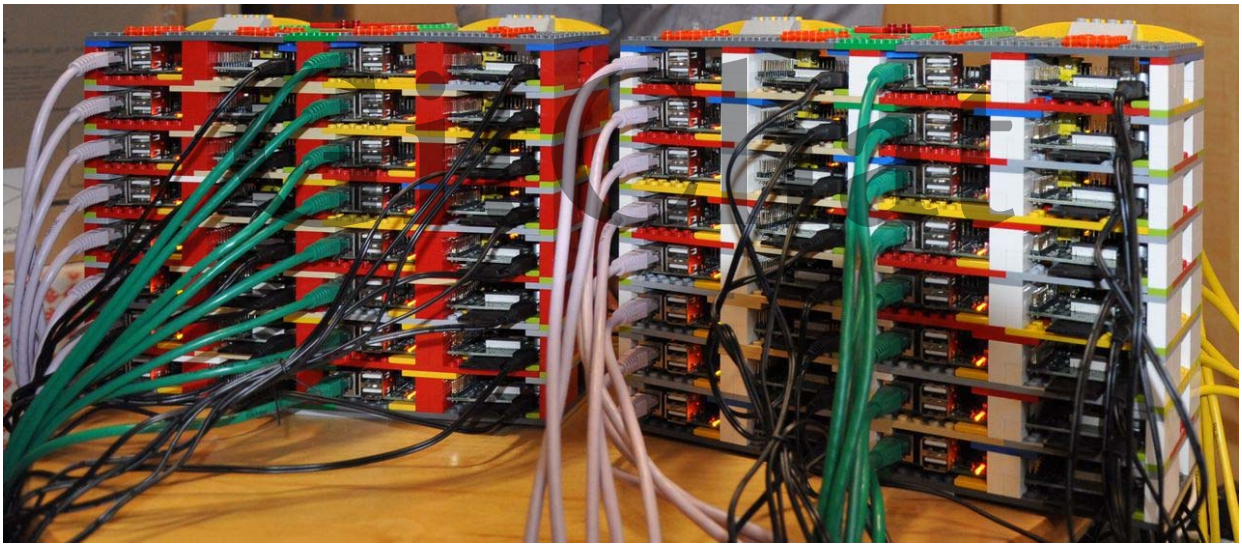
数据，在大厂这个问题能很好解决，若没在，就需要自己到处收集数据和爬数据。多添置些硬件、一有空就开启爬虫、收集开放数据。通过这些数据去实践数据挖掘、机器学习等实验。

算力，有了编程语言、算法、数据等软资产外，我们还需要更多更强的硬件设备。在大厂这个问题也迎刃而解，若是个人和小团队，就自己Diy机器、攒设备了。

同时，我们发起了一个开源项目：Super Compute Project

本项目意在将超级计算平民化，程序员、初创团队都可拥有自己的小型计算集群和桌面超级计算机，拥有自己的算力。

你可尝试搭建一个树莓派超级计算机，子弹(钱)多些的，可搭建一个Mac Mini集群(堆叠个10台Mini，家里的电源应该可以支撑)。



钱再宽裕的，就DIy自己的GPU、FPGA集群。

一些技术参考，大家可访问灰狐百科资源索引。

https://github.com/huihoo/wiki

这里不好意思哈，我贴了很多灰狐Wiki地址，最主要是我们都在时时更新它们，希望对大家有帮助，也渴望和大家能一起协作。

三、学会运营：社交网络、增长黑客、数据挖掘

社交网络、邮件营销、磁力营销等病毒营销相关概念，连同搜索引擎优化、众包、协作等，共同构成了“黑客增长术”的概念。

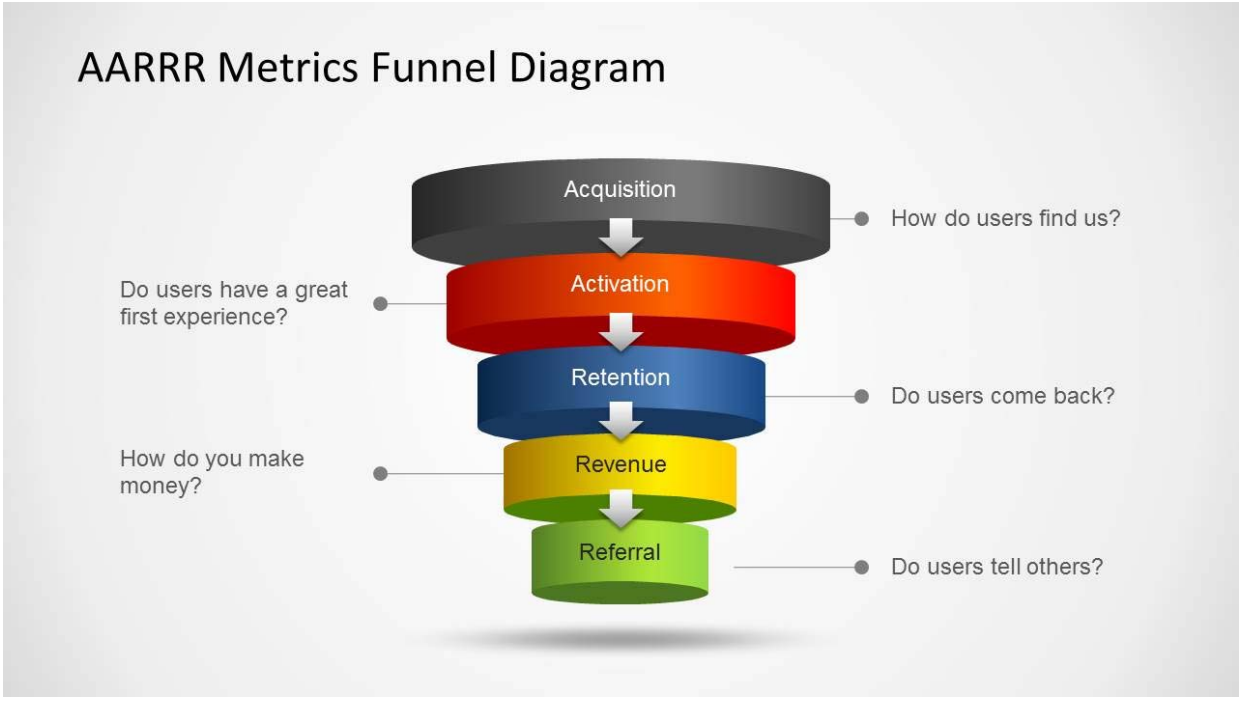
是否懂得运用黑客增长术和如何研发产品服务将同等重要。

合格的增长黑客需要有跨多个学科的积淀

图片来自：https://www.biaodianfu.com/growth-hacking.html

类似这样的定义还有很多，简单讲增长黑客是个跨界人才。是游走在产品、运营、研发、设计、用研等环节间的多面手，是介于极客、发明家和广告狂人之间的角色，是个复合型人才。

增长黑客要干什么事？增长目标给出了答案：



（图片来自：slidemodel.com）

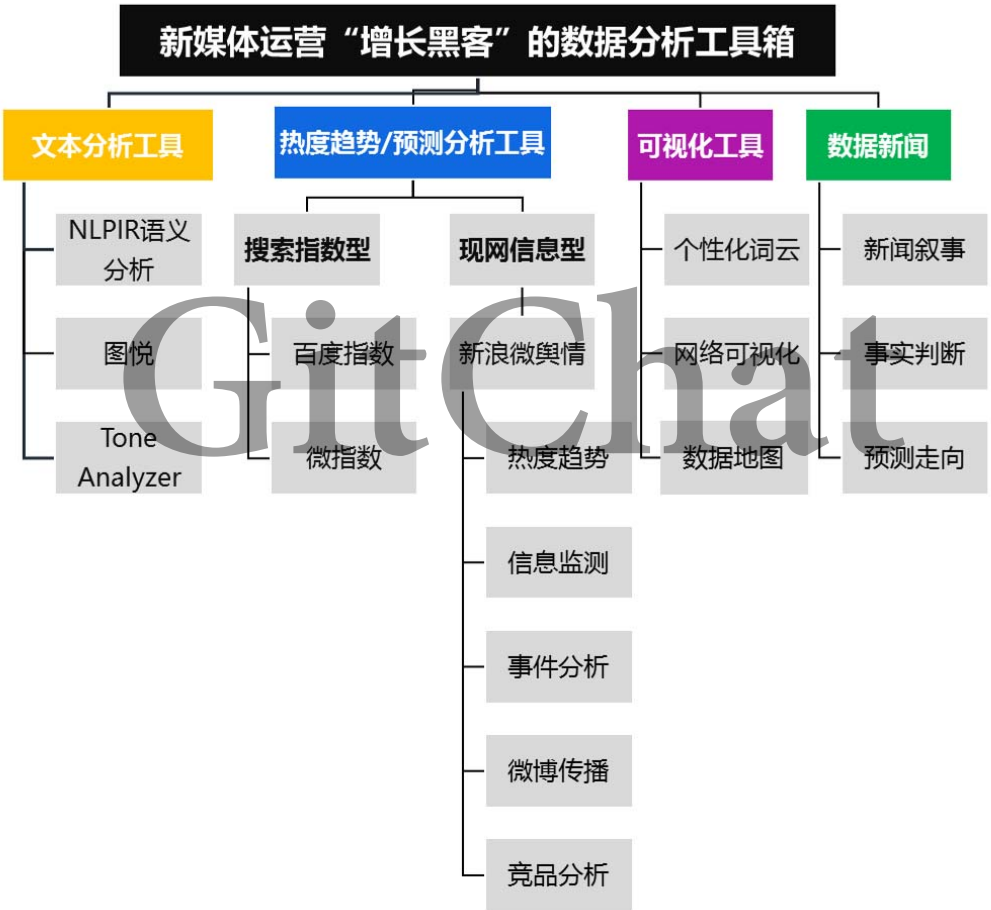
AARRR转化漏斗模型：Acquisition（获取用户）、Activation（激发活跃）、Retention（提高留存）、Revenue（增加收入）、Referral（传播推荐）

大家可依照这5个环节，进行一次次实操，逐渐形成自己的最佳实践，创建属于自己的增长模型。

有了这些模型目标，我们需要借助一些工具帮助我们达成目标。

因为一切用数据说话，所以增长黑客最核心的就是数据分析工具。

工具有很多，这里推荐新媒体运营的“增长黑客”数据分析工具箱



知乎上数据分析工具的讨论：

https://www.zhihu.com/topic/19569775/hot

从我个人的角度来看，NLP是数据挖掘最直接和广泛的应用范畴，也是你进入人工智能领域一个非常好的切入点，它所需要的各类成本也相对较低(如硬件等)，而且我们每天接触最多的也是各种媒体内容，当然现在视频内容也非常多，所以计算机视觉你也可更多关注。

参考：《社交网站的数据挖掘与分析》

有了目标和工具，接下来就是开干。我推荐的几个数据分析方向，供大家参考：

1. 开放数据，各类公开的开放数据集。

   https://github.com/caesar0301/awesome-public-datasets

   https://github.com/okulbilisim/awesome-datascience#data-sets

2. 社交网络

   有大量的社交网路数据可获得，且网络上有大量的实战可供大家参考。

3. 区块链和数字货币

   区块链和比特币一直很火，且数据都是公开的，所以做数据挖掘和分析是比较方便和有价值的。

4. 电竞数据

   Dota2

   https://developer.valvesoftware.com

   https://github.com/kronusme/dota2-api

英雄联盟的数据分析

- https://developer.riotgames.com/
- https://github.com/pseudonym117/Riot-Watcher
- https://github.com/meraki-analytics/cassiopeia
- https://github.com/simoncos/lola/

除了自己使用和搭建数据分析环境外，也可借助外部的SaaS服务快速切入增长黑客领域。

- GrowingIO为产品和运营打造的数据分析服务。
- 网易七鱼 以云客服为核心，较为传统。

四、模仿大牛：自由软件和开源圈是技术大牛们出没的地方, GitHub，技术会议

GitHub和各类开源基金会聚焦了众多的技术大牛，去找找他们，技术会议让你有机会近距离和大牛交流。

知乎这几年发展迅猛，大量牛人、大咖纷纷入驻，直接关注他们。知乎的话题质量很高，是对维基百科的深度补充。

我关注的部分技术牛人，供参考：

- https://github.com/dsyme F#之父
- https://github.com/jboner Akka之父
- https://github.com/nathanmarz Apache Storm创始人
- https://github.com/psyeugenic Erlang开发者
- https://github.com/tqchen 陈天奇，MXNet开发者
- https://mli.github.io/ 李沐，MXNet开发者
- https://github.com/Unknwon Gogs作者，Go语言牛人

  …

另外，有关增长黑客，知乎上可关注他们：

- 《增长黑客》作者 @范冰XDash
- Facebook 数据分析专家 @邹昕
- GrowingIO CEO @张溪梦 Simon 张老师是前 LinkedIn 商业分析部门高级总监。

五、享受人生：自由、开放、协作、分享

自由、开放、协作、分享

这是灰狐的发展理念，这个理念很多年前就作为我们行事的准则，如何更好达成，我们也在不断实践和探索中。

自由

现实世界，自由不易。对我们来讲，获得自由意味着更多。

我们希望自己和更多人都能过上自由的生活。

虽然，现实很残酷，但理想不变。

我们会不断践行各种自由形态下的自由工作、自由学习、自由生活。

开放

开放心态、开放业务和基础架构、抱团取暖。大家想想看，若没有类似OpenStack云平台和Hadoop大数据这样的开源平台，我们普通人和一般公司是很难切入这些领域的，单独一家公司也是几乎不可能开发出这样规模的软件堆栈的。

对人和事，我们都需要保持开放包容的心态，不轻易否定一个人和一件事，去融合平衡好周围的人和事。

这里有我6年前写得一篇博客《从开源到开放，新的商业模式》供大家参考。

协作

这是个协作的时代，我们崇拜天才、英雄，但现在已不再是一个人的时代了。
所有的商业和组织都在寻找高效的协作方式，因为协作正在改变世界。

分享

生命即为分享 Life is for sharing 我们相信人的天性是乐于分享的 - Share and Enjoy!

共享经济其实也可以简单理解为一种分享经济。

好了，就先分享到这，谢谢大家。接下来，期待与你的更多交流与协作。