数据分析有三大问：

- 如何入门数据分析？
- 数据分析有哪些工具？
- 怎么做数据分析？

关于第一问，想必读完我公众号《10 周入门系列文章》的同学，应该有大致的认识。

今天开始讲第二大问题——数据分析工具！

对于数据分析，我一直强调核心是业务，通过业务的分析逻辑影射到数据分析的处理逻辑，而数据分析工具则是帮助我们实现结果的手段。

但是，你说工具不重要吧，他又很重要，就像什么样的路选择什么样的交通工具，合适的工具能帮我们更快的达到终点。对应数据分析的不同环节，也要选择不同的工具，甚至选择更容易上手。

今天这篇文章，就是来扫盲工具的。

> PS：估计网上没有比这个更全面的了

一、从工具属性和分析师需求来划分

在企业中，数据分析师往往分为业务和技术两类，两者能力和工作内容有较大区别，对于工具的要求也各有侧重。

业务 or 技术

业务类分析师，往往在营运部，市场部，销售部等，根据服务的业务部门的不同，可能叫数据运营，经营分析，会员分析，商业分析师等名字。因为各个业务线具体考虑的问题不同，分析思路与体系均

有不同，所以会有这种区别。日常的工作更多是整理业务报表，针对特定业务做专题分析，围绕业务增长做需要用到数据的测算、规划、方案等。

技术类分析师，往往在 IT 部、数据中心。根据从事的工作环节不同，被分成数据库工程师，ETL 工程师，爬虫工程师，算法工程师等角色。在中小企业，往往一个技术小哥通吃这些流程。在大企业，一个标准的数据中心，一般都有数据仓库、专题分析、建模分析等组来完成数据开发工作，再大的公司，还有专门负责数据治理的小组。之所以有这个区分，是因为生产数据，需要一个多层次的复杂的数据系统。一个数据系统，需要数据采集、数据集成、数据库管理、数据算法开发、报表设计几个环节组合。这样才能把分散在各处的一点一滴的数据集中起来，计算成常用的指标，展示成各种炫酷的图表。这里每一个环节都需要对应的技术支持和人员工作，因此有了不同的岗位。

> PS：大家在找数据分析岗时，一定要区分是技术还是业务，和自己的职业倾向是否匹配。

分析师有技术和业务之分，那对应工具也有这样的属性侧重。

分析类工具

对于初级数据分析师，玩转 Excel 是必须的，数据透视表和公式使用必须熟练，VBA 是加分。另外，还要学会一个统计分析工具，SPSS 作为入门是比较好的。

对于高级数据分析师，使用分析工具是核心能力，VBA 基本必备，SPSS/SAS/R 至少要熟练使用其中之一，其他分析工具（如 Matlab）视情况而定。

对于数据挖掘工程师……嗯，R 和 Python 必备，要靠写代码来解决。

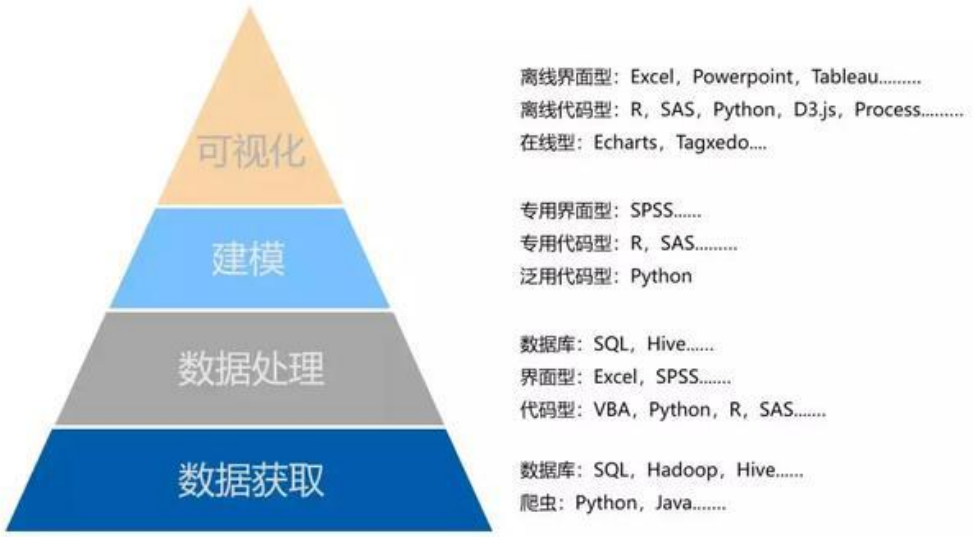
代码类工具

对于初级数据分析师，会写 SQL 查询，有需要的话写写 Hadoop 和 Hive 查询，基本就 OK 了。

对于高级数据分析师，除了 SQL 以外，学习 Python 是很有必要的，用来获取和处理数据都是事半功倍。当然其他编程语言也是可以的。

对于数据挖掘工程师，Hadoop 得熟悉，Python/Java/C++至少得熟悉一门，Shell 得会用……总之编程语言绝对是数据挖掘工程师的最核心能力。

一图说明问题：



二、从企业数据应用架构来划分

工具的使用还要看企业的需求和环境。为什么小企业招数据分析师其实就是 Excel 做报表，大企业找数据分析是却是把玩高大上的 Python、R？这就要看企业的数据架构。

站在 IT 的角度，实际应用中可以把数据工具分为两个维度：

第一维度：数据存储层——数据报表层——数据分析层——数据展现层

第二维度：用户级——部门级——企业级——BI 级

1、数据存储层

数据存储设计到数据库的概念和数据库语言，这方面不一定要深钻研，毕竟有专业的 DBA。但至少要理解数据的存储方式，数据的基本结构和数据类型。SQL 查询语言必不可少，精通最好。可从常用的 selece 查询，update 修改，delete 删除，insert 插入的基本结构和读取入手。

Access 这是最基本的个人数据库，经常用于个人或部分基本的数据存储；MySQL 数据库，这个对于部门级或者互联网的数据库应用是必要的，这个时候关键掌握数据库的库结构和 SQL 语言的数据查询能力。SQL Server2005 或更高版本，对中小企业，一些大型企业也可以采用 SQL Server 数据库，其实这个时候本身除了数据存储，也包括了数据报表和数据分析了。

DB2，Oracle 数据库都是大型数据库，主要是企业级，特别是大型企业或者对数据海量存储需求的就是必须的了，一般大型数据库公司都提供非常好的数据整合应用平台。

BI 级别，实际上这个不是数据库，而是建立在前面数据库基础上的，企业级应用的数据仓库。Data Warehouse，建立在 DW 机上的数据存储基本上都是商业智能平台，整合了各种数据分析，报表、分析和展现。

2、报表/BI 层

企业存储了数据需要读取，需要展现，报表工具则是最普遍应用的工具，尤其是在国内。过去传统报表大多解决的是展现问题，如今衍生了一些分析型报表工具，也会和其他应用交叉，做数据分析报表，通过接口开放功能、填报、决策报表功能，能够做到打通数据的进出，涵盖了早期商业智能的功能。

像 Tableau、PowerBI、FineBI、Qlikview 这类 BI（商业智能）工具，涵盖了报表、数据分析、可视化等多层。底层还可于数据仓库衔接，构建 OLAP 分析模型。

3、数据分析层

这个层其实有很多分析工具，当然我们最常用的就是 Excel。

Excel 软件，首先版本越高越好用这是肯定的。当然对 excel 来讲很多人只是掌握了 5%Excel 功能，Excel 功能非常强大，甚至可以完成所有的统计分析工作！但是我也常说，有能力把 Excel 玩成统计工具不如专门学会统计软件。

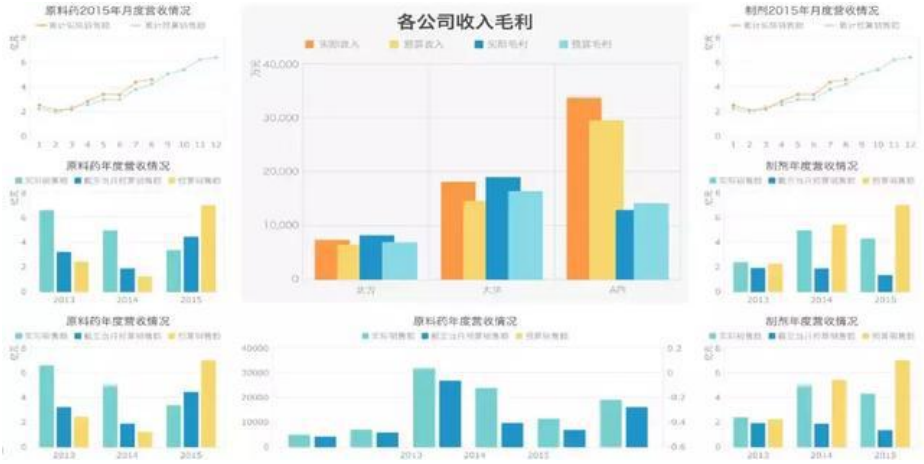
SPSS 软件：当前版本是 18，名字也改成了 PASW Statistics；我从 3.0 开始 Dos 环境下编程分析，到现在版本的变迁也可以看出 SPSS 社会科学统计软件包的变化，从重视医学、化学等开始越来越重视商业分析，现在已经成为了预测分析软件。

SAS 软件：SAS 相对 SPSS 其实功能更强大，SAS 是平台化的，EM 挖掘模块平台整合，相对来讲，SAS 比较难学些，但如果掌握了 SAS 会更有价值，比如离散选择模型，抽样问题，正交实验设计等还是 SAS 比较好用，另外，SAS 的学习材料比较多。

其他还有 Python 和 R，后面还会详细讲。

4、表现层

表现层也叫数据可视化，以上每种工具都几乎提供了一点展现功能。但要说企业级最常应用的还是 BI，做分析做报告。



> PS：需要说明的是，这样的分类并不是区分软件，只是想说明软件的应用。有时候我们把数据库就用来进行报表分析 ,有时候报表就是分析 ,有时候分析就是展现；当然有时候展现就是分析，分析也是报表，报表就是数据存储了！

三、4 大工具盘点

以上啰嗦了那么多，具体讲讲 Excel、R、Python、BI 吧。

>>>>Excel

适用场景：

1.一般的办公需求下的数据处理工作；

2.中小公司数据管理，存储（很多国有企业都用）；

3.学校学生，老师做简单的统计分析（如方差分析，回归分析）；

4.结合 Word，PowerPoint 制作数据分析报告；

5.数据分析师的主力分析工具（部分数据分析师的辅助工具）；

6.部分商业杂志，报刊图表制作（数据可视化）；

优点：

1.容易上手；

2.学习资源十分丰富；

3.可以用 Excel 做很多事情，建模，可视化，报表，动态图表；

4.帮助你在进一步学习其它工具之前（比如 Python，R），理解很多操作的含义；

缺点：

1.深入学习需要掌握 VBA，难度还是很高；

2.当数据量较大时，会出现卡顿的情况；

3.到 Excel2016 版 ,在不借助其它工具的情况下 ,Excel 数据文件本身能够容纳的数据仅有 108 万行，不适合处理大规模数据集；

4.内置统计分析种类太简单，实用价值不大；

5.不像 Python，R 语言等开源软件，正版 Excel 需要付费，比如我用 office365.每年需要支付 300 多块钱（不过也值了）

>>>>R

使用场景：

通过扩展的第三方 R 包，R 能够做的事情几乎涵盖了任何需要数据的领域。就我们一般的数据分析或者学术数据分析工作而言，R 能做的事情包括但不限于如下方面：

1.数据清洗与整理；

2.网络爬虫；

3.数据可视化；

4.统计假设检验（t 检验，方差分析，卡方检验等）；

5.统计建模（线性回归，逻辑回归，树模型，神经网络等）；

6.数据分析报告输出（Rmarkdown）；

R 容易学吗？

从我个人来看，想要入门 R 是非常简单的，10 天的集中学习，对于掌握 R 的基本使用，基本数据结构，数据导入导出，简单的数据可视化，是完全没有问题的。有了这些基础，在遇到实际的问题时，去找到需要使用的 R 包，通过阅读 R 的帮助文档，以及网络上的资料，就能够相对快速的解决具体问题了。

>>>>Python

R 语言和 Python 同为需要编程的数据分析工具，所不同的是，R 专门用于数据分析领域，而科学计算与数据分析只是 Python 的一个应用分支，Python 还可以用来开发 web 页面，开发游戏，做系统的后端开发，以及运维工作。

现在的一个趋势是，Python 在数据分析领域正在追赶 R，在某些方面已经超越了 R，比如机器学习，文本挖掘等偏编程的领域，但 R 语言在偏统计的领域仍然保持优势。Python 在数据分析方面的发展，很多地方借鉴了 R 语言中的一些特色。所以，如果你现在还是一片空白，还没开始学习，要做决定学习 R 还是 Python 的话，建议从 Python 入手。

Python 和 R 都比较容易学习，但是如果你同时学习两者，由于在很多地方它们非常相似，就会很容易混淆，所以建议不要同时学习它们。等其中一个掌握到一定的程度，再着手学习另外一个。

Python 能做什么？

1.网络数据爬取，使用 Python 能够很容易的编写强大的爬虫，抓取网络数据；

2.数据清洗；

3.数据建模；

4.根据业务场景和实际问题构造数据分析算法；

5.数据可视化（个人感觉不如 R 好用）；

6.机器学习，文本挖掘等高级数据挖掘与分析领域；

应该学习 R 还是 Python？

如果因为时间有限，只能选择其中的一种来学习的话，我建议使用 Python。但我仍然建议两者都了解一下，毕竟每个人都不一样。可能你在某些地方听说，Python 在工作中更加常用，但是工作中，解决问题才是最重要的，如果你能够用 R 高效的解决问题，那就用 R。实际上，Python 很多数据分析方面的特色，是模仿 R 来实现的，比如 pandas 的数据框，正在开发中的 ggplot 可视化包模仿的是 R 语言中非常著名的 ggplot2.

>>>>BI

多数分析师日常的工作就是做报表，而数据分析师更多用到的报表是 BI。

BI 全称商业智能，在传统企业中，它是一套完整的解决方案。将企业的数据有效整合，快速制作出报表以作出决策。涉及数据仓库，ETL，OLAP，权限控制等模块。

BI 工具主要有两种用途。一种是利用 BI 制作自动化报表，数据类工作每天都会接触大量数据，并且需要整理汇总，这是一块很大的工作量。这部分工作可以交给 BI 自动化完成，从数据规整、建模到下载。

另外一种是使用其可视化功能进行分析，BI 的优点在于它提供比 Excel 更丰富的可视化功能，操作简单上手，而且美观，如果大家每天作图需要两小时，BI 会缩短一半时间。

BI 作为企业级应用，可以通过它连接公司数据库，实现企业级报表的制作。这块涉及数据架构，就不深入讲了。

关于 BI，像 Tableau、PowerBI、FineBI、Qlikview 这类 BI（商业智能）工具，涵盖了报表、数据分析、可视化等多层。底层还可于数据仓库衔接，构建 OLAP 分析模型。

个人觉得，要想快速上手数据分析，前期数据思维的养成，BI 工具无疑是最容易上手的。下一篇文章，就要教大家动手搭建 BI 分析平台，并学会操作一款 BI 工具！