04 | 如何挑选适合项目场景的数据分析工具?

1. 数据分析整体流程

- 1.1. 明确问题
- 1.2. 搭建框架
- 1.3. 数据提取
- 1.4. 数据处理
- 1.5. 数据分析
- 1.6. 数据展现
- 1.7. 撰写报告
- 1.8. 报告演讲
- 1.9. 报告闭环

2. 常用工具、场景对比

工具	应用场景	掌握程度
Mysql、hive	基本上所有的数据获取方式 可以进一步学习一些Linux命令	超级熟练 数据提取不能出错
Excel	最高频、最有机会展示的数据处理工具 举例：老板让你现场画个图	超级熟练
R	统计语言，就是为数据分析而生，简单易学，但计算能力较差 举例：2G数据导入可能就死机	熟练
Python	脚本工具，可扩展性极强，算法研发同学必备，数据分析以Pands包为主，其他常用包含爬虫，文本挖掘	熟练

3. Excel 常用操作

- 3.1. Excel 对比分析（筛选和色阶功能）
- 3.2. Excel—时间序列拆解分析（透视图功能）
- 3.3. Excel—相关性分析（常用函数功能）
- 3.4. Excel—临界点分析（插入图表复杂功能）

4. SQL 常见问题

- 4.1. 常见现象：
  - 有同学一旦表关联较多，内部逻辑稍微复杂，就怀疑自己的代码准确性
- 4.2. 解决方案：
  - 若公司内部有SQL高手，请教、模仿他的方法
  - 若公司内大家水平差不多，就靠自己，可刷Leetcode题练习
- 4.3. 判断SQL熟练与否的标准：
  - 随时让你跑个数，都能自信弄出来！
- 4.4. 常见问题：
  - Max函数问题
  - 日期处理问题
  - 聚合计数问题
  - 一列变多行问题
  - 取TOP问题
  - 数据倾斜问题

5. R 语言以及 Python 脚本案例

- 5.1. R语言/Python常见问题：
  - 把业务问题转化为机器语言，进而用代码实现，最后帮业务找到解决问题的切入点！
- 5.2. Python相比R的更多价值：
  - 爬虫 — 请注意界限，不要触碰法律红线！
  - 文本挖掘 — 对评论运营很有帮助
  - UDF 函数 — Hive里面用Python自定义UDF函数可很快解决问题
  - 对于打算从事算法研发同学 — 个性化推荐、底层运维、Web 开发等