# 学习心得

——20大数据2班高嘉仪

通过这次张鹏飞老师主讲的实训，让我对Hadoop有了更深的了解。

Hadoop是一个由Apache基金会所开发的[分布式系统](https://baike.baidu.com/item/%E5%88%86%E5%B8%83%E5%BC%8F%E7%B3%BB%E7%BB%9F/4905336?fromModule=lemma_inlink" \t "https://baike.baidu.com/item/Hadoop/_blank)基础架构。用户可以在不了解分布式底层细节的情况下，开发分布式程序。充分利用集群的威力进行高速运算和存储。HDFS有高[容错性](https://baike.baidu.com/item/%E5%AE%B9%E9%94%99%E6%80%A7/9131391?fromModule=lemma_inlink" \t "https://baike.baidu.com/item/Hadoop/_blank)的特点，并且设计用来部署在低廉的硬件上；而且它提供高吞吐量来访问[应用程序](https://baike.baidu.com/item/%E5%BA%94%E7%94%A8%E7%A8%8B%E5%BA%8F/5985445?fromModule=lemma_inlink" \t "https://baike.baidu.com/item/Hadoop/_blank)的数据，适合那些有着超大数据集的应用程序。Hadoop还有高可靠性、高扩展性、效性、高容错性、低成本五大优点，所有被人们广泛使用。

大数据必然无法用单台的计算机进行处理，必须采用分布式架构。它的特色在于对海量数据进行分布式数据挖掘。但它必须依托云计算的分布式处理、分布式数据库和云存储、虚拟化技术。Linux操作系统具有稳定性、安全性、资源消耗小等优点，是搭建分布式集群的最佳选择。

大数据的基础在于海量数据的云存储，如何获取海量数据，并对其进行分布式存储是大数据的首要问题。在本次学习过程中我们学会了在项目上埋点，并且直接在hdfs集群中生成数据；还学会了使用hive对数据实现需求文档设计以及将结果落地到MySQL，实现了后端的所有工作；如何将所得的数据更好地呈现在页面上，我们可以利用superset或echarts在前端实现数据可视化。

大数据的意义不在于对海量数据的存储，而在于对海量数据的计算。归根结底在于四个字：“辅助决策”，运用大数据分析总结经验、发现规律、预测趋势。因此可知数据本身不产省价值，大数据必须和其他具体的领域、行业结合，能够给企业决策提供帮助后，才具有价值。很多企业和政府部门等机构都可以借助大数据，提升管理、决策水平。

对大数据这门在IT界广泛应用的课程而言，我的能力范围仅限于老师在视频中所讲解的知识点，要想成为一名大数据研发工程师还需要付出很多时间和精力去学习更深度的知识并且探索其中的奥秘。