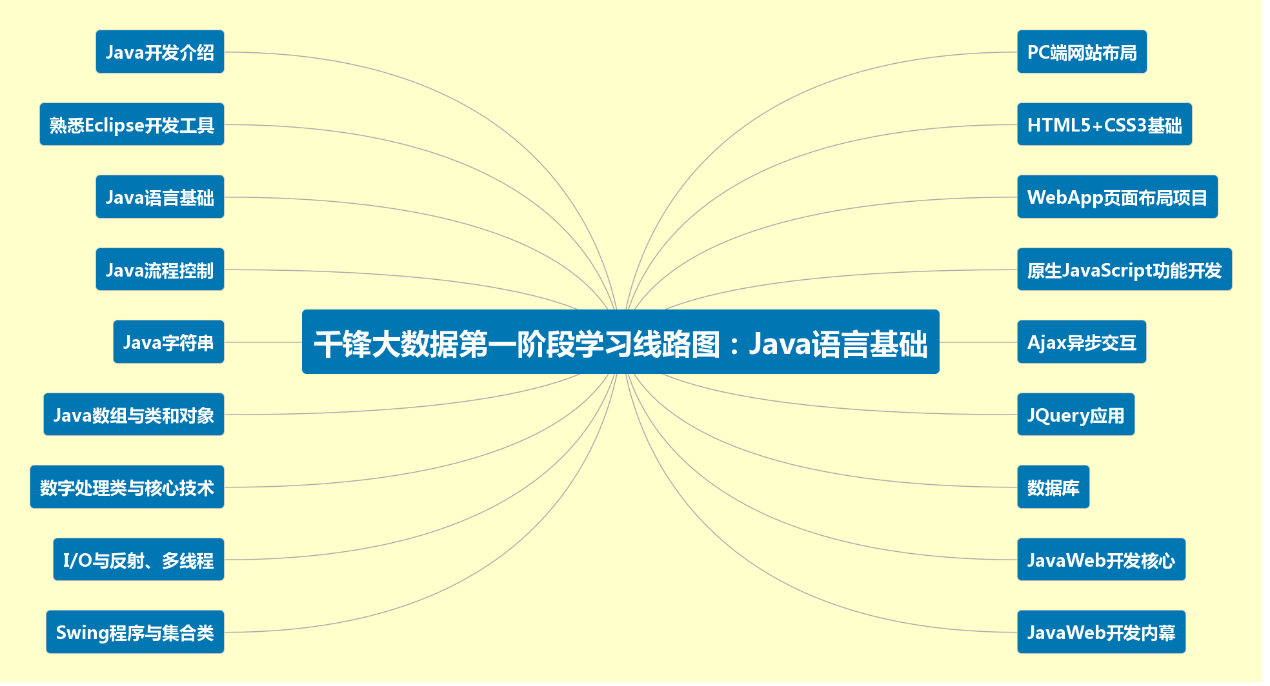
# **大数据的学习历程**

**学习大数据的头一步：打好基础，比高逼格的大数据技术更重要！而基础知识就是：编程语言和linux操作系统。**

我们以java编程为例，当然了你所擅长的编程语言也可以是python、Scala等！



**Java**：只需要学习Java的标准版JavaSE就可以了，像Servlet、JSP、Tomcat、Struts、Spring、Hibernate，Mybatis都是JavaEE方向的技术在大数据技术里用到的并不多，只需要了解就可以了，当然Java怎么连接数据库还是要知道的，像JDBC一定要掌握一下，有同学说Hibernate或Mybites也能连接数据库啊，为什么不学习一下，我这里不是说学这些不好，而是说学这些可能会用你很多时间，工作中也不常用，我还没看到谁做大数据处理用到这两个东西的，当然你的精力很充足的话，可以学学Hibernate或Mybites的原理，不要只学API，这样可以增加你对Java操作数据库的理解，因为这两个技术的核心就是Java的反射加上JDBC的各种使用。

**Linux：**因为大数据相关软件都是在Linux上运行的，所以Linux要学习的扎实一些，学好Linux对你快速掌握大数据相关技术会有很大的帮助，能让你更好的理解hadoop、hive、hbase、spark等大数据软件的运行环境和网络环境配置，能少踩很多坑，学会shell就能看懂脚本这样能更容易理解和配置大数据集群。还能让你对以后新出的大数据技术学习起来更快。



**接下来就是大数据技术的两大生态系统：Hadoop生态系统和spark生态系统。**

**Hadoop：**这是现在流行的大数据处理平台几乎已经成为大数据的代名词，所以这个是必学的。Hadoop里面包括几个组件HDFS、MapReduce和YARN，HDFS是存储数据的地方就像我们电脑的硬盘一样文件都存储在这个上面，MapReduce是对数据进行处理计算的，它有个特点就是不管多大的数据只要给它时间它就能把数据跑完，但是时间可能不是很快所以它叫数据的批处理。YARN是体现Hadoop平台概念的重要组件有了它大数据生态体系的其它软件就能在hadoop上运行了，这样就能更好的利用HDFS大存储的优势和节省更多的资源比如我们就不用再单独建一个spark的集群了，让它直接跑在现有的hadoop yarn上面就可以了。



**Spark：**它是用来弥补基于MapReduce处理数据速度上的缺点，它的特点是把数据装载到内存中计算而不是去读慢的要死进化还特别慢的硬盘。特别适合做迭代运算，所以算法流们特别稀饭它。它是用scala编写的。Java语言或者Scala都可以操作它，因为它们都是用JVM的。

**以上很多都是必须的理论技术知识，想要成功就业，光纸上谈兵不行，还需要一定的实战经验才行，寻找一些大数据相关的项目进行练练手，巩固一下自己的技术。**



以上就是对如何学习大数据的简单总结，如果学完之后仍有余力，自然也是可以学习机器学习、人工智能等技术知识吗，对你以后的就业会有很大的优势！