**大数据项目常见开发步骤**

**一 大数据项目开发步骤**

**（1）步骤**

第一步：需求：数据的输入和数据的产出；

第二步：数据量、处理效率、可靠性、可维护性、简洁性；

第三步：数据建模；

第四步：架构设计：数据怎么进来，输出怎么展示，最最重要的是处理流出数据的架构；

第五步：再次思考大数据系统和企业IT系统的交互；

第六步：最终确定选择、规范等；

第七步：基于数据建模写基础服务代码；

第八步：正式编写第一个模块；

第九步：实现其它的模块，并完成测试和调试等；

第十步：测试和验收；

**（2）大数据流程**

从流程角度上看，整个大数据处理可分成4个主要步骤。

第一步是数据的搜集与存储;

第二步是通过数据分析技术对数据进行探索性研究，包括无关数据的剔除，即数据清洗，与寻找数据的模式探索数据的价值所在;

第三步为在基本数据分析的基础上，选择和开发数据分析算法，对数据进行建模。从数据中提取有价值的信息，这其实是真正的阿里云大数据的学习过程。这当中会涉及很多算法和技术，比如机器学习算法等;

最后一步是对模型的部署和应用，即把研究出来的模型应用到生产环境之中。

1) 数据采集：定制开发采集程序，或使用开源框架flume

2) 数据预处理：定制开发mapreduce程序运行于hadoop集群

3) 数据仓库技术：基于hadoop之上的Hive

4) 数据计算基数：基于spark计算框架

5) 数据导出：基于hadoop的sqoop数据导入导出工具

6) 数据可视化：定制开发web程序或使用superset、kettle等产品