基于Hadoop&Spark的关联规则实践计划

大数据

——曾楠嵘

1. 实践目的

> 通过前一段时间的知识学习，我对Hadoop、Saprk的基本框架、运行模式等有了大致的了解。制定本次实践计划，其主要目的在于，希望通过亲身实践，加深自己对Hadoop、Spark这两类大数据工具的理解，熟悉其从部署到运作的基本操作流程，了解相关关联规则算法原理，掌握算法的基本应用，为日后的大数据学习与实践积累经验。

2. 实践计划

实践主要分为**前期模拟**，与**后期实践。**其中前期模拟在**个人笔记本**中进行，后期实践在**INTEL NUC**中进行。下面为初步制定的实践流程安排。

2.1 前期模拟

- [x] Linux服务集群搭建(基于VMware)
  - 四台Centos7服务器，一台用作Master,其余三台用作workers

- [x] 相关软件安装
  - [x] JDK安装（版本 1.8）
  - [x] Hadoop安装（版本 3.1.1）
  - [x] Sacla安装（版本 2.12.8）
  - [x] Spark安装（版本 2.4.0）

- [ ] 相关理论知识学习
  - [ ] Hadoop相关
    - [x] HDFS架构
    - [x] MapReduce基本流程
    - [ ] Yarn容器管理
  - [ ] Spark相关

- [x] Scala基本框架
- [x] spark运行模式（standalone、yarn）
- [ ] RDD基本原理了解

- [ ] 关联规则相关
  - [ ] FPGrowth算法原理

- [x] 模拟实践
  - [x] HDFS文件管理

2.2 后期实践

- [ ] Linux服务集群搭建(基本KVM)
  - 8台, 一台用作NameNode, 一台用作SecondaryNameNode,其余六台用作workers

- [ ] 模拟实践
  - [ ] 于spark-yarn模式下运行SparkPi，观察效率，并模拟故障情景
  - [ ] 于spark-yarn模式下运行Groceries(啤酒尿布案例)，分析结果

2.3 时间安排

- 对于剩余的理论知识学习部分，即Yarn容器管理，RDD基本原理，FPGrowth算法原理，预计可以在 1月4号 完成（期间除去针对 1月3号 操作系统 考试 的两天复习）
- 对于后期实践部分，即**NUC的部署与模拟实践**，以及最终的**实践总结**，预计可以在 1月14 左右完成（期间除去两门已知的 JAVAEE考试 与 物联网组网考试 ,具体时间可能会根据剩下的一门 数值分析考试 以及 传感器实验 时间进行1~2天的浮动)