BD_第 02 周作业_吴超_心得

这堂课作为本次课程的第二节课，让我对大数据仓库原理与搭建有了宏观的认识。和老师一起从“0”开始搭建大数据架构设备，为今后在集群上进行分布式 AI 大数据分析和机器学习，研发奠定了良好的开发环境，硬件系统运维水平也有显著上升，非常的实用，受益匪浅。

关于硬件设备，我们学习了处理集群 Cluster 的搭建，虚拟机的使用，Ubantu Linux 系统的安装，广域和局域网络调试，安装包镜像源升级与下载等。

关于大数据组件，我们学习了 Zookeeper、Kafka、FileBeat、Mysql、Canal 的基本原理，软件安装与配置，组件启动方法。

关于大数据理论，首先，我学习到了系统中，常见数据的类型：Web Action Event、Database、DiskFile 等。其次，了解了数据采集 ETL 阶段，大数据的常用传输链路。然后，学习到了两种数据变化的监控和捕捉方法：A-磁盘文件→Filebeat→Logstash→Kafka；B-数据库→Canal Server→Kafka。

在大数据仓库搭建实操作业中，我的配置方案如下：

1. 环境配置-安装 VirtualBox 虚拟机，模拟硬件节点设备环境。
2. 系统配置-在虚拟机中，安装 Ubantu Linux 系统。
3. 辅助配置-安装 VirtualBox 加强工具和扩展包，实现屏幕自适应、虚拟机与宿主机之间的粘贴板共享和文件夹共享。

4. 网络配置-在该虚拟机中第一块网卡中采用 NAT 模式，实现广域网连接和 Internet 浏览；在该虚拟中第二块网卡中采用 Host-Only 模式，实现局域网连接和机器间通信。
5. 集群配置-在第一台机器完成配置后，用 Clone 的方式复制若干台子节点，修改 hostname、Ip、MAC 等信息，从单机拓展到集群环境，确保节点间可以 Ping 通，并远程访问。
6. 互信配置-在超级用户和普通用户下分别生成秘钥对，通过公钥的复制和共享，实现机器间的免密登录，提高效率。
7. 时钟配置-利用 ntpd 服务器定时(1 分钟)同步机器时钟，确保机群系统时钟一致，所有数据的生成/修改时间相同。
8. 同步脚本配置-编写 msh 脚本，实现所有节点间命令同步操作，如安装软件、查看时钟、查看硬件使用情况等。避免手动在多节点间复制命令，提高效率。脚本执行方法：msh bigdata-node 1 3 df –h。
9. 拷贝脚本配置-编写 xsync 脚本，实现所有节点间相同路径下文件拷贝，脚本执行方法：xsync 文件绝对路径。
10. 软件包配置-解压老师给下载好的大数据软件包，存储到指定路径下。并通过 xsync 脚本分发给所有处理节点。
11. 软件包安装-先后顺序为，Zookeeper，Kafka，Kafka-Manager，FileBeat，MySql，Canal。

在本次课程中，搭建了大数据的基本架构环境，为后续的学习和试验奠定了硬件设备基础，也提升了我的系统运维水平。