2020/6/22

Hadoop 是什么

1. Hadoop 就是存储海量数据和分析海量数据的工具。
2. Hadoop 是由 java 语言编写的，在分布式服务器集群上存储海量数据并运行分布式分析应用 的开源框架，其核心部件是 HDFS 与 MapReduce。
3. HDFS 是一个分布式文件系统：引入存放文件元数据信息的服务器 Namenode 和实际存放数 据的服务器 Datanode，对数据进行分布式储存和读取。
4. MapReduce 是一个计算框架：MapReduce 的核心思想是把计算任务分配给集群内的服务 器里执行。通过对计算任务的拆分（Map 计算/Reduce 计算）再根据任务调度器 （JobTracker）对任务进行分布式计算。
5. HDFS 为海量的数据提供了存储，则 MapReduce 为海量的数据提供了计算。把 HDFS 理解为 一个分布式的，有冗余备份的，可以动态扩展的用来存储大规模数据的大硬盘。把 MapReduce 理解成为一个计算引擎，按照 MapReduce 的规则编写 Map 计算/Reduce 计算的 程序，可以完成计算任务。

使用 Hadoop

1、搭建 Hadoop 集群 无论是在 windows 上装几台虚拟机玩 Hadoop，还是真实的服务器来玩，说简单点就是把 Hadoop 的安装包放在每一台服务器上，改改配置，启动就完成了 Hadoop 集群的搭建。

2、上传文件到 Hadoop 集群 Hadoop 集群搭建好以后，可以通过 web页面查看集群的情况，还可以通过 Hadoop 命令来 上传文件到 hdfs 集群，通过 Hadoop 命令在 hdfs 集群上建立目录，通过 Hadoop 命令删除集 群上的文件等等。

3、编写 map/reduce 程序 通过集成开发工具（例如 eclipse）导入Hadoop 相关的 jar 包，编写 map/reduce 程序，将程 序打成 jar 包扔在集群上执行，运行后出

计算结果。