|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 《大数据技术原理与应用》实验报告 | | | | |
| 题目： | Hadoop安装配置 | 姓名 |  | 日期: |
| 实验环境：   1. 采用WSL2（Windows Subsystem for Linux，Windows的GNU/Linux子系统），因为WSL2能有调用NVCC和CUDA的GPU。 | | | | |
| 实验内容与完成情况：  以前（所以，此处时间并不是最近）就使用过VMware还有物理机安装各种发行版本的GNU/Linux。    还有现在我VMware中其他虚拟机，    也干过自己魔改Ubuntu玩，和同学（自动化的，我转专业过来，但是还住在自动化）们暑假进行了Gnu/Linux系统界面美化大赛。  所以，这次我将采用WSL2（Windows Subsystem for Linux，Windows的GNU/Linux子系统）方式搭建Hadoop环境。  先用wsl --install -d ubuntu-22.04拉去最新稳定版的镜像并安装，同时开启sshd。      接着我从（https://download.oracle.com/otn/java/jdk/8u391-b13/b291ca3e0c8548b5a51d5a5f50063037/jdk-8u391-linux-x64.tar.gz?AuthParam=1698928962\_190d2d1e8836492fbbc220a84e4d71ee）下载包并安装jdk，    之后我用wget从（hadoop-3.3.6.tar.gz (tsinghua.edu.cn)）下载Hadoop包    具有jdk的三个实例已创建，  配置/etc/hosts文件，  ​  已经打通ssh并用Xshell成功连接，    现在将之前下载好的hadoop包解压并配置~/.bashrc环境变量，    为了便于后期各节点之间的连接，所以此处添加SSH免密码公钥连接，    此处是公钥生成，    在slave1得到一个为下文的公钥：  The key's randomart image is:  +---[RSA 3072]----+  | . .o+.o . .|  | o.o . \* + |  | .. + \* |  |. o. . . o |  | o o + S o |  |o o \* o. . |  |+o.+ B o |  |+=oEB.o . |  |.o\*+oB+ |  +----[SHA256]-----+  同理分别得到所以node的免密公钥。  接着开始配置master实例的hadoop-3.3.6/etc/hadoop/core-site.xml ，    以及hdfs-site.xml、yarn-site.xml、mapred-site.xml、works文件等,          紧接着进行格式文件系统hdfs namenode -format、start-dfs.sh（hdfs）、start-yarn.sh（yarn）、start-all.sh（全部），    最终经过各种折腾，测试Hadoop在web端运行结果。 | | | | |
| 出现的问题：  1. 国内访问Ubuntu官方镜像源不稳定。    2.IP地址每次开机变化。  3.出现以下与公钥和密码有关提示，    4.Hadoop2.x与3.x之间有一定差距，导致不能照搬群内的教程 | | | | |
| 解决方案（列出遇到的问题和解决办法，列出没有解决的问题）：  1.对/etc/apt/source.list进行换源为国内的源，例如中科大源、兰大源等，提高速率。  以下我采用了省内的兰州大学镜像源      2.将DHCP修改改静态IP地址，还有通过配置Hyper-V管理器，    3.将公钥挨个进行配置，    4.以下是2.x与3.x主要区别，自行在网络上找寻的。  # hadoop 2.x和hadoop 3.x的功能比较  这部分内容将让你知道22个hadoop2和hadoop3最大的区别，让我们来一个个讨论吧。  ## 1. 证书  - Hadoop 2.x - Apache 2.0，Open Source  - Hadoop 3.x - Apache 2.0，Open Source  ## 2. Java的最小支持版本  - Hadoop 2.x - Java 7  - Hadoop 3.x - Java 8  ## 3. 容错性  - Hadoop 2.x - 通过版本复制提高容错性  - Hadoop 3.x - 通过[擦除编码技术](https://data-flair.training/blogs/hadoop-hdfs-erasure-coding/)实现容错性，这个特性我会再写一篇技术文章分析它怎么实现，以及有哪些优势  ## 4. 数据负载均衡  - Hadoop 2.x - 对于数据，采用hdfs均衡器做负载均衡  - Hadoop 3.x - 对于数据，采用数据内节点均衡器做负载均衡，通过HDFS磁盘均衡器CLI调用  ## 5. 存储方案  - Hadoop 2.x - 3倍副本存储方案  - Hadoop 3.x - 在HDFS上支持擦除编码技术（Erasure Coding，简称EC），解决副本复制和副本存储带来的空间和资源的消耗，以EC代替副本，提供和副本存储相同的容错能力，并且存储开销不大于单副本开销的50%。  ## 6. 存储开销  - Hadoop 2.x - 2倍额外的存储空间开销  - Hadoop 3.x - 0.5倍额外的存储空间开销  ## 7. 存储开销实例  - Hadoop 2.x - 如果有6个block块，那么基于副本方案，将会有18个块空间被使用  - Hadoop 3.x - 如果有6个block块，那么基于EC技术方案，将会有9个块空间被使用，6个数据块和3个奇偶校验块  ## 8. YARN时间线服务  - Hadoop 2.x - 使用老的时间线服务，有性能扩展问题  - Hadoop 3.x - 升级时间线服务v2，提高了稳定性和扩展性  ## 9. 默认的端口区间  - Hadoop 2.x - 在hadoop2.x中，一些默认端口是linux的临时端口范围，当服务启动时，可能无法绑定造成启动失败  - Hadoop 3.x - 3.x中，这些端口从临时端口范围中移出来了  ## 10. 工具  - Hadoop 2.x - 使用Hive Pig Tez Hama Giraph和其他hadoop工具  - Hadoop 3.x - Hive Pig Tez Hama Giraph和其他hadoop工具都可以使用  ## 11. 兼容的文件系统  - Hadoop 2.x - HDFS，FTP文件系统（将所有数据存储在远端FTP服务器上），亚马逊S3，微软Azure WASB文件系统  - Hadoop 3.x - 以上及微软Azure数据湖文件系统  ## 12. 数据节点资源  - Hadoop 2.x - DataNode节点资源并非专用与MapReduce任务，也可以将其用于其他应用程序  - Hadoop 3.x - DataNode节点资源也可以使用到其他应用程序  ## 13. MR API兼容性  - Hadoop 2.x - Hadoop 1.x应用程序可以兼容2.x  - Hadoop 3.x - Hadoop 1.x应用程序可以兼容到3.x中  ## 14. 是否支持windows  - Hadoop 2.x - 支持  - Hadoop 3.x - 支持  ## 15. 插槽、容器  - Hadoop 2.x - Hadoop 1任务运行在slot的概念中，hadoop 2任务运行在容器container概念中  - Hadoop 3.x - Hadoop 3任务运行在容器概念中  ## 16. 单点故障  - Hadoop 2.x - 当namenode下线时，standby namenode自动上线提供服务，支持namednode单点故障恢复  - Hadoop 3.x - 支持namednode单点故障恢复  ## 17. HDFS联盟  - Hadoop 2.x - Hadoop 1中，仅一个NameNode管理多个namespace命名空间，Hadoop 2中，多个NameNode管理多个命名空间  - Hadoop 3.x - Hadoop 3中，多个NameNode管理多个命名空间  ## 18. 扩展性  - Hadoop 2.x - 在Hadoop 2中，单个集群可以扩展到10000个节点  - Hadoop 3.x - 扩展性更好，单个集群可以扩展到大于10000个节点  ## 19. 更快地访问数据  - Hadoop 2.x - 由于DataNode有缓存机制，我们可以快速访问到数据  - Hadoop 3.x - DataNode也有缓存机制  ## 20. HDFS快照  - Hadoop 2.x - Hadoop 2添加了快照的支持，当用户错误时提供了数据的灾难恢复和保护机制  - Hadoop 3.x - 也提供了快照支持  ## 21. 平台  - Hadoop 2.x - 可以作为一个平台，运行事件处理、流计算、实时计算等不同类型数据分析  - Hadoop 3.x - 也可以作为一个平台，在YARN上运行上述处理任务  ## 22. 集群资源管理  - Hadoop 2.x - 使用YARN作为集群资源管理，提高扩展性、高可用、多租户  - Hadoop 3.x - 使用YARN进行资源管理  # 总结  综上所述，我们讨论了Hadoop 2.x和hadoop 3.x的比较，现在我们可以根据自己的情况做出选择，到底是2.x好还是3.x好。  这意思内容导致它们的配置文件并不是完全相同。 | | | | |