**山东大学 计算机科学与技术 学院**

**云计算技术 课程实验报告**

|  |  |  |  |
| --- | --- | --- | --- |
| **学号：**201900130133 | **姓名：**施政良 | | **班级：**四班 |
| **实验题目：**面向分布式存储和计算的Hadoop练习 | | | |
| **实验学时：**2 | | **实验日期：**2020-04-18 | |
| **实验目的：**在Linux环境下，熟悉Hadoop环境。  **具体包括：**了解Hadoop生态结构和关键工具/组件，了解Hadoop编程环境的配置和部署，完成实验环境及实验工具的熟悉，撰写实验报告。 | | | |
| **硬件环境：**  联网计算机一台 | | | |
| **软件环境：**  Windows or Linux | | | |
| **实验步骤与内容：**  **实验步骤概述：**  本次试验旨在介绍Hadoop的基本概念，需要了解Hadoop生态结构和关键工具/组件。同时需要熟悉Linux下Hadoop的开发环境，完成实验环境及实验工具的配置。  **具体实验内容**   1. Hadoop介绍   Apache Hadoop是一款支持数据密集型分布式应用程序并以Apache 2.0许可协议发布的开源软体框架。它支持在商用硬件构建的大型集群上运行的应用程序。Hadoop是根据谷歌公司发表的MapReduce和Google档案系统的论文自行实作而成。所有的Hadoop模块都有一个基本假设，即硬件故障是常见情况，应该由框架自动处理。  Hadoop框架透明地为应用提供可靠性和数据移动。它实现了名为MapReduce的编程范式：应用程序被分割成许多小部分，而每个部分都能在集群中的任意节点上执行或重新执行。此外，Hadoop还提供了分布式文件系统，用以存储所有计算节点的数据，这为整个集群带来了非常高的带宽。MapReduce和分布式文件系统的设计，使得整个框架能够自动处理节点故障。它使应用程序与成千上万的独立计算的电脑和PB级的数据连接起来。现在普遍认为整个Apache Hadoop“平台”包括Hadoop内核、MapReduce、Hadoop分布式文件系统（HDFS）以及一些相关项目，有Apache Hive和Apache HBase等等  2. Hadoop的组成架构  Hadoop 由许多元素构成。其最底部是 Hadoop Distributed File System（HDFS），它存储 Hadoop 集群中所有存储节点上的文件。HDFS的上一层是MapReduce 引擎，该引擎由 JobTrackers 和 TaskTrackers 组成。通过对Hadoop分布式计算平台最核心的分布式文件系统HDFS、MapReduce处理过程，以及数据仓库工具Hive和分布式数据库Hbase的介绍，基本涵盖了Hadoop分布式平台的所有技术核心  3. Hadoop环境配置  3.1 安装对应版本的虚拟机  由于在之前的实验中使用Ubuntu18.04作为实验环境，因此在本次试验中首先安装Ubuntu20.04作为Linux实验环境。  同实验一，在本次实验中采用 VMware 虚拟机，并采用 ubuntu 配置 linux 环境。Ubuntu 一个 以桌面应用为主的 Linux 操作系统其界面与常用的 Windows 界面非常相似。解决了 Linux 安装和使用困难的问题，同时，其通过 sudo 指令执行系统相关的任务的设置也 使其比传统的以系统管理员账号进行管理工作的方式更为安全。  打开 vmware 并创建虚拟机    将 ubuntu 镜像文件装入虚拟机中并选择合适的硬件配置（如磁盘大小，内核数量） 即可。最终完成虚拟机的创建，如下图所示。    3.2创建 hadoop 用户  在终端中输入如下命令创建Hadoop用户   1. 为 hadoop 用户设置密码：sudo passwd hadoop 2. 为 hadoop 用户增加管理员权限：sudo adduser hadoop sudo # 3. 注销当前用户，并使用 hadoop 用户登录：su - hadoop   3.3 安装 JDK  JDK为java运行的必要环境。在本次实验中，首选创建 java 目录，例如#创建 jvm 文件夹  mkdir /usr/lib/jvm  解压到目录下  sudo tar zxvf jdk-18\_linux-aarch64\_bin.tar.gz -C /usr/lib/jvm  其中sudo为root权限，tar为解压命令，zxvf为tar的命令行参数，jdk-18\_linux-aarch64\_bin.tar.gz是jdk压缩包的文件名  上述指令将JDK解压到/usr/lib/jvm 目录下。  之后进入该目录  cd /usr/lib/jvm  为了便于后续实验的进行，此处使用mv指令，将文件夹重命名为java文件名规范。  3.4配置 java 环境变量  在终端使用vim编辑器对根目录下的./bashrc文件进行编辑，添加相应的文件路径  vim ~/.bashrc  在~/.bashrc 最后添加下列代码并保存   |  | | --- | | #Java Environment  export JAVA\_HOME=/usr/lib/jvm/java  export JRE\_HOME=${JAVA\_HOME}/jre  export CLASSPATH=.:${JAVA\_HOME}/lib:${JRE\_HOME}/lib  export PATH=${JAVA\_HOME}/bin:$PATH |   **3.5更新配置，并测试是否安装成功**  配置java环境之后，需要更新并测试安装是否成功。使用如下命令使新配置的环境变量生效  source ~/.bashrc  打印Java版本，测是否安装成功  java -version  如下图所示    **3.5.** **安装 ssh**  首先安装 SSH server  sudo apt-get install openssh-server  之后登录本机测试，需要手动输入“yes”  ssh localhost  过程如下所示    **3.6 安装单机 Hadoop**  解压安装 Hadoop到/usr/local 目录下  之后进入目录  **cd /usr/local**  配置环境变量，使用vim在~/.bashrc 中添加如下代码并保存   |  | | --- | | #Hadoop Environment  export HADOOP\_HOME=/usr/local/hadoop  export CLASSPATH=$($HADOOP\_HOME/bin/hadoop classpath):$CLASSPATH  export HADOOP\_COMMON\_LIB\_NATIVE\_DIR=$HADOOP\_HOME/lib/native  export PATH=$PATH:$HADOOP\_HOME/bin:$HADOOP\_HOME/sbin |   刷新~/.bashrc 配置文件  source ./bashrc  之后输入 hadoop version，测试是否安装成功    **3.7 伪分布式 Hadoop**  在/usr/local/hadoop/etc/hadoop 目录 hadoop-env.sh 添加 Java 路径   |  | | --- | | export JAVA\_HOME=/usr/lib/jvm/java |   修改配置文件 core-site.xml Hadoop 的配置文件位于 /usr/local/hadoop/etc/hadoop/ 中修改 core-site.xml 文件，添加下列内容。   |  | | --- | | <configuration>  <property>  <name>hadoop.tmp.dir  </name>  <value> file:/usr/local/hadoop/tmp</value>  <description>Abase for other temporary directories.</description> </property>  <property>  <name>fs.defaultFS</name>  <value>hdfs://localhost:9000</value>  </property>  </configuration> |   修改配置文件 hdfs-site.xml添加下列内容，   |  | | --- | | <configuration>  <property>  <name>dfs.replication</name>  <value>1</value>  </property>  <property>  <name>dfs.namenode.name.dir</name>  <value>file:/usr/local/hadoop/tmp/dfs/name</value>  </property>  <property>  <name>dfs.datanode.data.dir</name>  <value>file:/usr/local/hadoop/tmp/dfs/data</value>  </property>  <property>  <name>dfs.http.address</name>  <value>0.0.0.0:50070</value>  </property>  </configuration> |   之后格式化集群节点 hdfs namenode -format，并启动 hadoop  start-dfs.sh    使用 jps 查看进程，出现 DataNode，NameNode,SecondaryNameNode 即启动成功。 在浏览器输入 localhost:50070 查看 hadoop 状态。    关闭 Hadoop stop-dfs.sh | | | |
| **结论分析与体会：**   1. **Hadoop的优点**   Hadoop是一个能够对大量数据进行分布式处理的软件框架。 Hadoop 以一种可靠、高效、可伸缩的方式进行数据处理。  Hadoop 是可靠的，因为它假设计算元素和存储会失败，因此它维护多个工作数据副本，确保能够针对失败的节点重新分布处理。  Hadoop 是高效的，因为它以并行的方式工作，通过并行处理加快处理速度。  Hadoop 还是可伸缩的，能够处理 PB 级数据。此外，Hadoop 依赖于社区服务，因此它的成本比较低，任何人都可以使用。  Hadoop是一个能够让用户轻松架构和使用的分布式计算平台。用户可以轻松地在Hadoop上开发和运行处理海量数据的应用程序。  它主要有以下几个优点   * + 高可靠性。Hadoop按位存储和处理数据的能力值得人们信赖 。   + 高扩展性。Hadoop是在可用的计算机集簇间分配数据并完成计算任务的，这些集簇可以方便地扩展到数以千计的节点中 。   + 高效性。Hadoop能够在节点之间动态地移动数据，并保证各个节点的动态平衡，因此处理速度非常快   + 高容错性。Hadoop能够自动保存数据的多个副本，并且能够自动将失败的任务重新分配 。   + 低成本。与一体机、商用数据仓库以及QlikView、Yonghong Z-Suite等数据集市相比，hadoop是开源的，项目的软件成本因此会大大降低 。   + Hadoop带有用Java语言编写的框架，因此运行在 Linux 生产平台上是非常理想的。Hadoop 上的应用程序也可以使用其他语言编写，比如 C++  1. **Hadoop有哪些应用**   Hadoop 的最常见用法之一是 Web 搜索。虽然它不是唯一的软件框架应用程序，但作为一个并行数据处理引擎，它的表现非常突出。Hadoop 最有趣的方面之一是 Map and Reduce 流程，它受到Google开发的启发。这个流程称为创建索引，它将 Web爬行器检索到的文本 Web 页面作为输入，并且将这些页面上的单词的频率报告作为结果。然后可以在整个 Web 搜索过程中使用这个结果从已定义的搜索参数中识别内容 | | | |