第二章 大数据处理框架Hadoop

第二章考点

- 分布式计算
- Hadoop内涵、特性、Hadoop 1.0和2.0的区别、Hadoop集群节点
- Linux基本指令
- HDFS基本操作

一、Hadoop简介

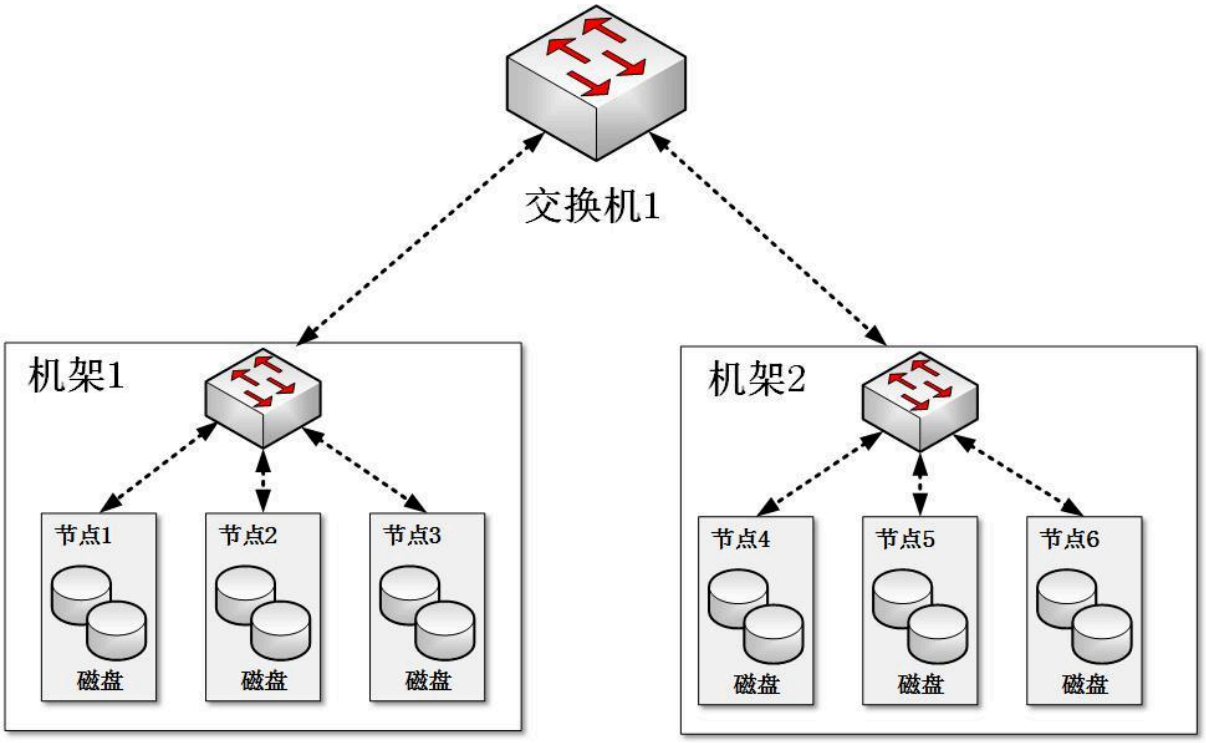
1. 分布式计算概述
   分布式计算是一种技术，它允许多台计算机协同工作以解决共同的问题
   分布式计算平台：是一种计算环境，它允许将一个大型的计算任务分解成多个子任务，由分布在不同位置的多个计算节点并行地进行计算
2. **Hadoop内涵：**
   Hadoop是一个开源的**分布式计算平台**，Hadoop是**基于JAVA开发**的，具有很好的**跨平台性**，支持Python，C/C++等编程语言
   Hadoop架构的核心是**分布式文件系统HDFS(数据存储)和MapReduce(数据处理)**
3. **Hadoop的特性**
   - **高可靠性**：采用冗余数据存储方式，即使一个副本发生故障，其他副本也可以正常运作
   - **高效性**：成百上千的服务器集中起来进行分布式存储和分布式处理
   - **高可扩展性**：成百上千台服务器进行分布式存储和分布式处理
   - **高容错性**：自动存储数据的多个副本，自动将失败的任务进行重新分配
   - **成本低**：普通用户很容易用PC搭建Hadoop运行环境；运行在Linux平台上，支持多种编程语言
4. Hadoop的应用现状
   在Hadoop于企业中的应用架构中：
   访问层（企业最典型的三种应用）：数据分析，数据实时查询，数据挖掘
   大数据层：离线分析，实时查询，BI分析
5. **Hadoop1.0和Hadoop2.0的区别**
   - MapReduce不再实现集群资源管理，并且引入YARN进行集群资源管理
   - 针对1.0中的HDFS存在NameNode单点故障问题，而Hadoop2.0通过引入NameNode HA(High Availability)解决了这一问题
   - 数据块大小：64MB $\rightarrow$ 128MB

二、Hadoop集群的部署与使用

1. 一个基本的**Hadoop集群中的节点**主要有：
   - NameNode：负责协调集群中的数据存储

- DataNode：存储被拆分的数据块
- JobTracker：协调数据计算任务
- TaskTracker：负责执行由JobTracker指派的任务
- SecondaryNameNode(冷备份)：帮助NameNode收集文件系统运行的状态信息

2. 集群网络拓扑
   - 普通的Hadoop集群结构由一个两阶段网络构成
   - 每个机架有30-40个服务器，配置一个1GB的交换机，并向上传输到一个核心交换机或者路由器(1GB或以上)



三、常见Linux命令

- `cd` 切换目录
  `cd /usr/local` 切换到指定路径目录
  `cd ..` 切换上一级目录
  `cd ~` 切换到用户自己的主文件夹
  `cd .` 切换到根目录
- `ls` 查看文件与目录
  `cd /usr`
  `ls -al` 查看当前路径下的所有文件和目录
- `mkdir` 新建目录
  `cd /temp`
  `mkdir a` 在 `/temp` 目录下创建目录
  `mkdir -p a1/a2/a3/a4` 其中 `-p` 的作用是确保中间必要的目录名称全部存在，如果不存在就再创建一个
- `cp` 复制文件或者目录
  `sudo cp ~/.bashrc /usr/bashrc1` 将 `.bashrc` 文件复制到目录 `/usr` 下并且重命名为 `bashrc1`

`cd /temp`
`mkdir test`
`sudo cp -r /temp/test/usr` 将创建的 `test` 目录复制到 `/usr` 目录下
其中 `-r` :表示递归处理，将指定目录文件和子目录一并处理

- `mv` 移动文件与目录，或者更名
  `sudo mv /usr/bashrc1 /usr/test` 将 `/usr` 目录下的文件bashrc1移动到 `/usr/test` 目录下
  `sudo mv /usr/test /usr/test2` 将 `test` 重命名为 `test2`
- `rm` 移除文件或者目录
  `sudo rm /usr/test2/bashrc` 将 `/usr/test2` 目录下的 `bashrc1` 文件删除
  `sudo rm -r /usr/test2` 将 `/usr` 目录下的 `test2` 目录删除
- `cat` 查看文件内容
  `cat ~/.bashrc`
- `tac` 反向查看文件内容
  `tac ~/.bashrc`
- `head` 取出前面几行
  `head -n 20 ~/.bashrc` 取出用户主文件夹下 `.bashrc` 内容前20行
  `head -n -50 ~/.bashrc` 查看当前用户文件夹下的 `.bashrc` 文件内容，后面50行不显示，只显示前面几行
- `tail` 取出后面几行
  `tail -n 20 ~/.bashrc` 取出后面20行
  `tail -n +50 ~/.bashrc` 列出50行以后的数据
- `touch` 修改文件时间或创建新文件
  `cd /tmp`
  `touch hello` 创建空文件 `hello`
  `ls -l hello` 查看当前文件时间
  `touch -d "5 days ago" hello` 将 `hello` 的时间修改问5天前
- `chown` 修改文件所有者权限
  `sudo chown root /tmp/hello`
  `ls -l /tmp/hello` 将 `hello` 文件所有者更换为 `root` 账号，并且查看属性
- `find` 文件查找
  `find ~ -name .bashrc` 在用户主文件夹中寻找文件名为 `.bashrc` 文件
- `tar` 压缩指令
  `sudo mkdir /test`
  `sudo tar -zcv -f /test.tar.gz test`
  在根目录 `/` 下新建文件夹 `test` 然后在根目录下打包成 `test.tar.gz`
  `sudo tar -zxv -f /test.tar.gz -C /tmp` 将上面的 `test.tar.gz` 解压到 `/tmp` 目录
- `grep` 查找字符串
  `grep -n 'examples' ~/.bashrc` 从 `~/.bashrc` 查找字符串 `examples`

四、常见Hadoop操作

1. 关于三种Shell命令方式的区别

- `hadoop fs` 适用于任何不同的文件系统，比如本地文件系统和HDFS文件系统
- `hadoop dfs` 只能适用于HDFS文件系统
- `hdfs dfs` 跟 `hadoop dfs` 的命令作用一样，也只能适用于HDFS文件系统

2. 使用hadoop用户登录Linux系统，启动Hadoop（Hadoop的安装目录为 “/usr/local/hadoop”），为hadoop用户在HDFS中创建用户目录 “/user/hadoop”

```
$ cd /usr/local/hadoop
$ ./sbin/start-dfs.sh #开启NameNode 和 DataNode 守护进程
$ ./bin/hdfs dfs -mkdir -p /user/hadoop
    ```
接着在HDFS的目录“/user/hadoop”下，创建test文件夹， 并查看文件列表
```bash
$ cd /usr/local/hadoop
$ ./bin/hdfs dfs -mkdir test
$ ./bin/hdfs dfs -ls .
```

将Linux系统本地的“~/.bashrc”文件上传到HDFS的test 文件夹中，并查看test

```
$ cd /usr/local/hadoop
$ ./bin/hdfs dfs -put ~/.bashrc test
$ ./bin/hdfs dfs -ls test
```

将HDFS文件夹test复制到Linux系统本地文件系统的 “/usr/local/hadoop”目录下

```
$ cd /usr/local/hadoop
$./bin/hdfs dfs -get test ./
```