1、Hadoop的安装和使用

- Hadoop简介
  - 基于Java语言开发，跨平台性好
  - 核心
    - 分布式文件系统HDFS
    - MapReduce
  - 版本
    - Hadoop 1.0
      - 0.20.x
        > 最后演化成1.0.x，变成了稳定版
      - 0.21.x
      - 0.22.x
    - 第二代Hadoop
      > 不同于Hadoop1.0，是一套全新的架构，均包含HDFS Federation和YARN
      - 0.23.x
      - 2.x
- 安装Hadoop前的准备工作
  - 创建hadoop用户
    > 使用hadoop用户登陆Linux系统，并为hadoop用户增加了管理员权限
  - 更新apt
  - 安装ssh
    - ssh
      > 是secure shell的缩写，是建立在应用层和传输层基础上的安全协议
    - ssh组成
      - 客户端
        > 包含ssh程序以及像scp，slogin以及sftp等其他的应用程序
      - 服务端
        > 是一个守护进程，在后台运行并响应来自客户端的连接请求
    - 安装ssh的原因
      > Hadoop名称节点需要启动集群中所有机器的Hadoop守护进程，这个过程就需要通过ssh登陆来实现。Hadoop并没有提供ssh输入密码登陆的形式，因此所有机器都需要配置为无密码登陆
  - 安装Java环境
- 安装Hadoop
  - 下载安装文件
    - 下载hadoop-2.7.1.tar.gz
    - 执行 sudo tar -zxf hadoop-2.7.1.tar.gz -C /usr/local
    - cd /usr/local
    - sudo mv ./hadoop-2.7.1/ ./hadoop
    - sudo chown -R hadoop ./hadoop

- 检查安装是否成果
  - cd /usr/local/hadoop
  - ./bin/hadoop version

- 单机模式配置

  Hadoop默认模式为非分布式（本地模式），无需进行其他配置即可运行。
- 伪分布式模式配置

  Hadoop可以在单个节点上以伪分布式的方式运行，同一个节点既作为名称节点又作为数据节点，读取的是分布式文件系统HDFS中的文件。需要指出的是，Hadoop的运行方式是由配置文件决定的，启动Hadoop时会读取配置文件，然后根据配置文件决定运行在什么模式下。
  - 修改配置文件

    配置文件位于/usr/local/hadoop/etc/hadoop/中
    - core-site.xml
      - 指定hadoop.tmp.dir

        用于保存临时文件
      - 指定fs.defaultFS

        指定HDFS的访问地址
    - hdfs-site.xml
      - dfs.replication

        指定副本的数量
      - dfs.namenode.name.dir

        用于设定名称节点的元数据保存目录
      - dfs.datanode.data.dir

        用于设定数据节点的数据保存目录
  - 执行名称节点格式化

    cd /usr/local/hadoop
    ./bin/hdfs namenode -format
  - 启动Hadoop

    cd /usr/local/hadoop
    ./sbin/start-dfs.sh
  - 使用web界面查看HDFS信息

    在地址栏输入http://localhost:50070就可以查看名称节点和数据节点信息
  - 运行Hadoop伪分布式实例
  - 关闭Hadoop

    /usr/local/hadoop/sbin/stop-dfs.sh
- 分布式模式配置

  HDFS的名称节点和数据节点位于不同机器上
  - 缺少设备，暂不安装
- 使用docker搭建Hadoop分布式集群
  - docker简介
  - 安装docker
  - 在docker上安装ubuntu系统
  - Ubuntu系统初始化（安装各种Hadoop需要的组建）
  - 安装Hadoop
  - 配置Hadoop集群
  - 运行Hadoop实例

幕布 - 思维概要整理工具