Hadoop

Java 开源 分布式基础架构

hadoop是一种分析和处理海量数据的**软件平台 (多个软件组成)**

Hadoop常用组件:

–HDFS: Hadoop分布式文件系统 (核心组件)

–MapReduce: 分布式计算框架 (核心组件)

–Y a r n: 集群资源管理系统 (核心组件)

Hadoop模式

–单机

–伪分布: 主要用来学习和测试

–完全分布

伪分布式的配置和完全分布式配置类似

1. HDFS角色

角色和概念

–Client

切分文件、访问HDFS、与NameNode交互,获取文件位置信息|读取和写入数据

–Namenode

头

–Secondarynode

小秘 定期合并fsimage和fsedits,推送给NameNode

–DataNode

工人 具体的存储点, 汇报存储信息给NameNode

概念

–名称空间和数据块的映射信息: namenode记录数据存哪了的”本”

–Fsimage: ”本”

–Fsedits: 文件变更日志

分割后一份128m

副本

配置文件:

–环境配置文件: hadoop-env.sh

–核心配置文件: core-site.xml

–HDFS配置文件: hdfs-site.xml

–节点配置文件: slaves

–

–

过程

• 储存和读取数据:

Client分割后---->问namenode存哪,存1,2,...盘并记录----->client自己存到1盘

Client----->namenode在哪,在1,2...盘并记录------>client自己找到1盘

• 修改数据

Fsedits: 文件变更日志

不改变源数据,打补丁方式

2.MapReduce角色

3.Yarn角色

################################################################

配置hadoop (192.168.1.60 namenode)

1. 安装openjdk **& java-1.8.0-openjdk-devel**

]# jps //验证

862 Jps //每敲一遍都会变

1. 安装hadoop

[root@nn01 ~]# cd hadoop/

[root@nn01 hadoop]# ls

hadoop-2.7.7.tar.gz kafka\_2.12-2.1.0.tgz zookeeper-3.4.13.tar.gz

解压hadoop包

[root@nn01 hadoop]# tar -xf hadoop-2.7.7.tar.gz

将包直接移动到习惯路径下

[root@nn01 hadoop]# mv hadoop-2.7.7 /usr/local/hadoop

[root@nn01 hadoop]# cd /usr/local/hadoop

[root@nn01 hadoop]# ls

工作目录:

bin include libexec NOTICE.txt sbin

etc lib LICENSE.txt README.txt share

[root@nn01 hadoop]# ./bin/hadoop //报错，JAVA\_HOME没有找到

解决:

[root@nn01 hadoop]# rpm -ql java-1.8.0-openjdk

第一行

[root@nn01 hadoop]# cd ./etc/hadoop/

修改配置文件

[root@nn01 hadoop]# vim hadoop-env.sh

25 export JAVA\_HOME="/usr/lib/jvm/java-1.8.0-openjdk-1.8.0.161-2.b14.el7.x86\_64/jre"

//java路径

33 export HADOOP\_CONF\_DIR="/usr/local/hadoop/etc/hadoop" //配置文件路径

使用:

[root@nn01 hadoop]# ./bin/hadoop jar \

> share/hadoop/mapreduce/hadoop-mapreduce-examples-2.7.7.jar \

> wordcount input output

//wordcount为参数 统计input这个文件夹，存到output这个文件里面

（这个文件不能存在，要是存在会报错，是为了防止数据覆盖）

#################################################################################

伪分布配置 (60 -> namenode , 192.168.1.61~63 -> datanode)

在nn01上配置好后,把工作目录发给其他节点

配置hosts

配置ssh信息关系(公钥,nn01传递其他节点,包括传给自己)

需修改的配置文件:

固定格式:

<property>

<name>.....</name>

<value>.....</value>

<description>.....</description>

</property>

环境配置文件: hadoop-env.sh

同上

核心配置文件: core-site.xml

<property>

<name>fs.defaultFS</name> //声明使用什么样的文件系统

<value>hdfs://nn01:9000</value> // “此内容” = “/”

</property>

<property>

<name>hadoop.tmp.dir</name>

<value>/var/hadoop</value>

</property>

HDFS配置文件: hdfs-site.xml

<property>

<name>dfs.namenode.http-address</name>

<value>nn01:50070</value>

</property>

<property>

<name>dfs.namenode.secondary.http-address</name>

<value>nn01:50090</value>

</property>

<property>

<name>dfs.replication</name>

<value>2</value>

</property>

节点配置文件: slaves

node1

node2

node3

初始化: ./bin/hdfs namenode -format

INFO

successfully formatted.

启动: ./sbin/start-dfs.sh //namenode远程连接其他机器启动服务

*//打jps可看到角色*

关闭: ./sbin/stop-dfs.sh

关闭会生成logs目录,可以排错用

Logs目录解析

-启动进程用户-

问题: 若datanode起不来,可能是数据不同步

解决:

关闭服务

rm -rf /var/hadoop/\*

rm -rf logs/\*

启动服务