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

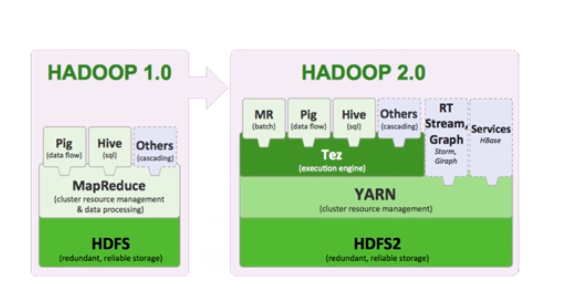
MapReduce:编程思想 Simplified Data Processing on Large Clusters

把一个某个复杂的计算任务分割成小的任务单元并行在各个节点上运行

搜集各个节点上运行结果合并运行，经过二次map，二次reduce，直到计算出结果位置；

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

一、Hadoop v2 架构图



二、Hadoop的运行模型

HDFS集群:

data\_node 数据存储节点 name\_node 名称节点 、secondary\_node辅助名称节点

YARN:集群资源管理

三、centos7安装Hadoop2.6.3

1.环境准备

centos7中一般已经自带JDK

[root@localhost zhanggen]# java -version

openjdk version "1.8.0\_102"

OpenJDK Runtime Environment (build 1.8.0\_102-b14)

OpenJDK 64-Bit Server VM (build 25.102-b14, mixed mode)

[root@localhost profile.d]# yum -y install java-1.8.0-openjdk\*

Centos7关闭防火墙

查看状态： systemctl status firewalld

开机禁用 ： systemctl disable firewalld

开机启用 ： systemctl enable firewalld

Centos7 关闭selinux服务

[root@localhost hdfs]# setenforce 1

[root@localhost hdfs]# getenforce

Enforcing

[root@localhost hdfs]# setenforce 0

[root@localhost hdfs]# getenforce

Permissive

1. 下载源码包

3.编译

[root@localhost bdapps]# mkdir /bdapps/

[root@localhost bdapps]# ls

hadoop-2.6.2

[root@localhost bdapps]# tar -zxvf /home/zhanggen/Desktop/hadoop-2.6.2.tar.gz -C /bdapps/

[root@localhost bdapps]# ln -sv /bdapps/hadoop-2.6.2 /bdapps/hadoop

‘/bdapps/hadoop’ -> ‘/bdapps/hadoop-2.6.2’

4.设置Java和Hadoop相关环境变量

source /etc/profile.d/hadoop.sh

source /etc/profile.d/java.sh

5.创建Hadoop用户

6.创建存储DataNode和 NameNode存储数据的目录

7.创建Hadoop存储日志的目录

yarn logs -applicationId application\_1551852706740\_0001 #查看任务执行日志

首先取到程序运行日志，再分析日志中出现错误的原因，解决问题

8.Hadoop主要配置文件（/bdapps/hadoop/etc/hadoop）

8.0.core-site.xml

针对NameNode IP地址 、端口（默认为8020）core-site.xml

8.1.hdfs-site.xml

针对HDFS相关的属性，每一个数据块的副本数量、NN和DA存储数据的目录 step6中创建的目录。

hdfs-site.xml

8.2.mapred-site.xml（指定使用yarn）

指定MapReduce是单独运行 还是运行在yarn之上，Hadoop2肯定是运行在yarn之上的；见

1. Hadoop的运行模型
2. mapred-site.xml

8.3.yarn-site.xml

yarn-site.xml 用于配置YARN进程及YARN的相关属性，首先需要指定ResourceManager守护进程的主机和监听的端口，对于伪分布式模型来讲，其主机为localhost，

默认的端口为8032；其次需要指定ResourceManager使用的scheduler，以及NodeManager的辅助服务。一个简要的配置示例如下所示：

vim yarn-site.xml

8.4.slave文件

slave文件存储了当前集群所有slave节点的列表，对于伪分布式模型，其文件内容仅应该为localhost，这特的确是这个文件的默认值。因此，为分布式模型中，次文件的内容保持默认即可。

如果服务器/虚拟机的进程起不来请确保本地 IP和配置文件里面的IP是否已经发生变化！

8.5.格式化HDFS

在HDFS的NN启动之前需要先初始化其用于存储数据的目录。如果hdfs-site.xml中dfs.namenode.name.dir属性指定的目录不存在，格式化命令会自动创建之；

如果事先存在，请确保其权限设置正确，此时格式操作会清除其内部的所有数据并重新建立一个新的文件系统，需要以hdfs用户的身份执行如下命令

复制代码

19/03/01 11:31:22 INFO namenode.FSImage: Allocated new BlockPoolId: BP-1276811871-127.0.0.1-1551411082356

19/03/01 11:31:22 INFO common.Storage: Storage directory /data/hadoop/hdfs/nn has been successfully formatted.

19/03/01 11:31:22 INFO namenode.NNStorageRetentionManager: Going to retain 1 images with txid >= 0

19/03/01 11:31:22 INFO util.ExitUtil: Exiting with status 0

19/03/01 11:31:22 INFO namenode.NameNode: SHUTDOWN\_MSG:

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

SHUTDOWN\_MSG: Shutting down NameNode at localhost/127.0.0.1

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

[hdfs@localhost hadoop]$ hdfs namenode -format

复制代码

9.启动Hadoop

HDFS格式化完成之后就可以启动 去/bdapps/hadoop/etc/hadoop目录下启动Hadoop的5大守护进程了。

9.1.启动HDFS集群

HDFS有3个守护进程：namenode、datanode和secondarynamenode，他们都表示通过hadoop-daemon.sh脚本启动或停止。以hadoop用户执行相关命令；

启动HDFS集群

HDFS集群web访问接口：

http://127.0.0.1:50070/dfshealth.html#tab-overview

9.2启动yarn集群

切换成yarn用户：YARN有2个守护进程：resourcemanager和nodemanager，它们通过yarn-daemon.sh脚本启动或者停止。以hadoop用户执行相关命令即可。

启动yarn集群

yarn集群web访问接口：

http://127.0.0.1:8088/cluster

10.测试

使用Hadoop自带的 hadoop-mapreduce-examples-2.6.2.jar，执行MapReduce任务是否可以正常执行，如果可以就意味着安装成功了。

在执行任务是要切换到hdfs用户下

执行任务

查看任务执行结果

11、python3调用HDFS集群API

Hadoop安装好了；（虽说是伪分布式的，如果要做分布式做好ssh免密码登录，把配置文件分发出去就好了）

但是我在网上看到python的pyhdfs模块可以调用HDFS集群的API进行上传、下载、查找....文件...于是储备下来了，也许可以用作后期 Hadoop自动化项目；

注意：在使用pyhdfs模块之前一定要确保Hadoop的配置文件都监听在外网端口并修改host文件。

1

192.168.226.142 localhost #windows hosts文件的路径 C:\WINDOWS\system32\drivers\etc\host Linux /etc/host

pip install pyhdfs -i http://pypi.douban.com/simple --trusted-host pypi.douban.com

使用pyhdfs模块调用HDFS集群API

四、Python3调用Hadoop MapReduce API

pip3 install mrjob -i http://pypi.douban.com/simple --trusted-host pypi.douban.com

hadoop fs -chown -R hadoop:hadoop /tmp #在执行MapReduce任务的时候hadoop用户会创建socket，通过jdbc访问。所以在执行你写得MapReduce之前一定要设置权限

MapReduce 任务工作流程（假设要对以下3行数据，统计词频）：

a b c

a c

a

第1步：map 把每个字符串映射成键、值对

(a,1)(b,1)(c,1)

(a,1)(c1)

(a1)

自动shuffle & sort：

shuffle: 把相同键的 值组合成1个的列表，（洗牌：平时玩牌的时候 把手里数字/字母相同的扑克牌们，码放在一起例如3A一起，两个2一起！）

sort: 再根据键排序；

(a,[1,1,1])

(b,[1])

(c,[1,1])

第2步：shuffle and sort之后，把键相同的值放到列表了，就方便reduce的时候对值进行计算、聚合操作（sum,mean,max）了！

(a,3)

(b,1)

(c,2)

python3版wordCount

python /MyMapReduce.py /a.txt -r hadoop #在Hadoop集群，执行Python的MapReduce任务。

执行结果

1.MapReduce案例

统计一下本周的报警情况

由于遗留了Zabbix报警未分类的问题，导致zabbix报警-----》转换到运维平台的工单信息---------》都是一个text字段！

本周报警信息

中国大学top10

求出得分前十的大学，和低消费 前十的大学

求唐县各乡/镇出现的最低、最高温度

MRStep连接多个mapper、reducer函数

求收入top3的行业

2.map + combine +reduce

map节点把所有集合计算的工作通过TCP协议传输到reduce节点会出现 单点负载压力的问题，所以combine出现了；

combine就是小的reduce，可以在map发送数据给reduce之前进行在map节点做初步的聚合运算，减小reduce节点的压力， 加速MapReduce任务的执行；