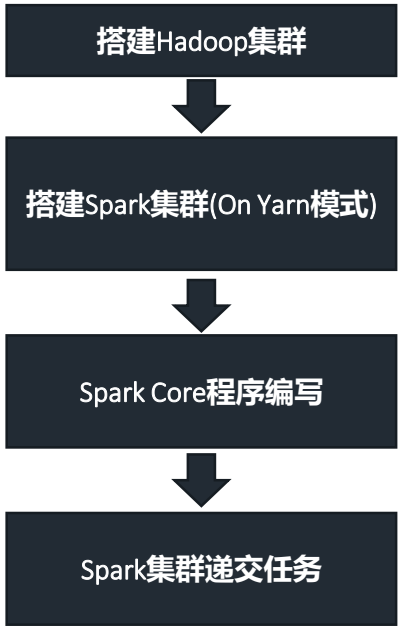
实验三： Spark单词计数

本次实验到底要做什么?

在Spark集群分布式的运行一个单词计数程序，并将结果输出到HDFS(Hadoop分布式文件系统)中。

你将收获：
(1)学会服务器购买与配置(如果你从事开发工作，一定会与linux服务器打交道)；
(2)掌握Hadoop、Spark集群的搭建方法(如果你从事大数据工作，一定离不开这两个集群)；
(3)使用Spark RDD处理数据(弹性分布式数据集处理数据的流程)；
(4)了解大数据的处理流程。(从大数据存储到大数据处理的实战)



目录

CONTENTS

搭建Hadoop集群

- 如何配置
- 启动与停止

如何配置

你需要参照实验手册，完成华为云3台服务器的购买。

windows使用putty或华为云提供的shell；mac和linux直接在终端使用ssh命令即可
ssh root@ip;

使用vim更改hosts文件，使得当前主机知道哪些主机名对应哪些ip：

```
192.168.0.138 master
192.168.0.245 slave01
192.168.0.61 slave02
```

由于Hadoop是java进程,所以需要添加jdk来运行；解压后，添加环境变量即可；

start-all.sh启动集群；
Jps查看java进程状态；

登录节点 → 更改主机名 → 主机名查询表 → 配置互信节点 → 安装jdk → Hadoop集群部署 → 启动集群

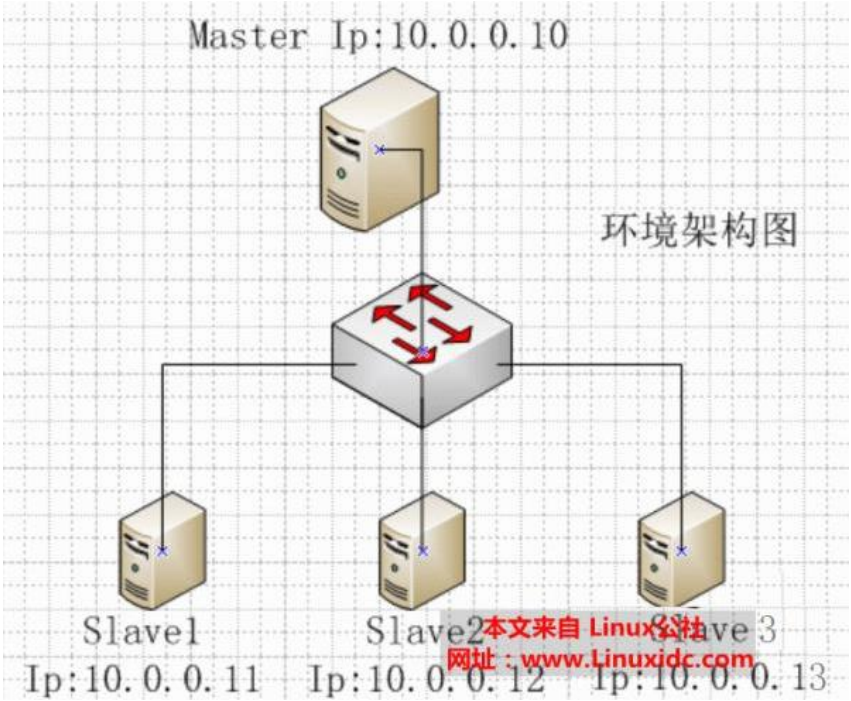
三个服务器分别命名为：master、slave01、slave02；重启后会更新命令提示符，如：

```
[root@master ~]#
```

使得master可以免密登录slave01和02，从而master可以管理slave01和slave02节点；其实就是给master添加slave01和slave02的登录密码；

1)解压Hadoop压缩包；2)进入etc/hadoop目录，修改配置文件：core-site.xml(指定namenode地址;使用hadoop时产生文件的存放目录等)；hdfs-site.xml(副本数量、hdfs权限设置等)；yarn-site.xml(yarn地址、对外暴露的接口等)；配置节点目录文件；3)将Hadoop解压包发送给节点；4)配置环境变量；

为什么要这样配置?



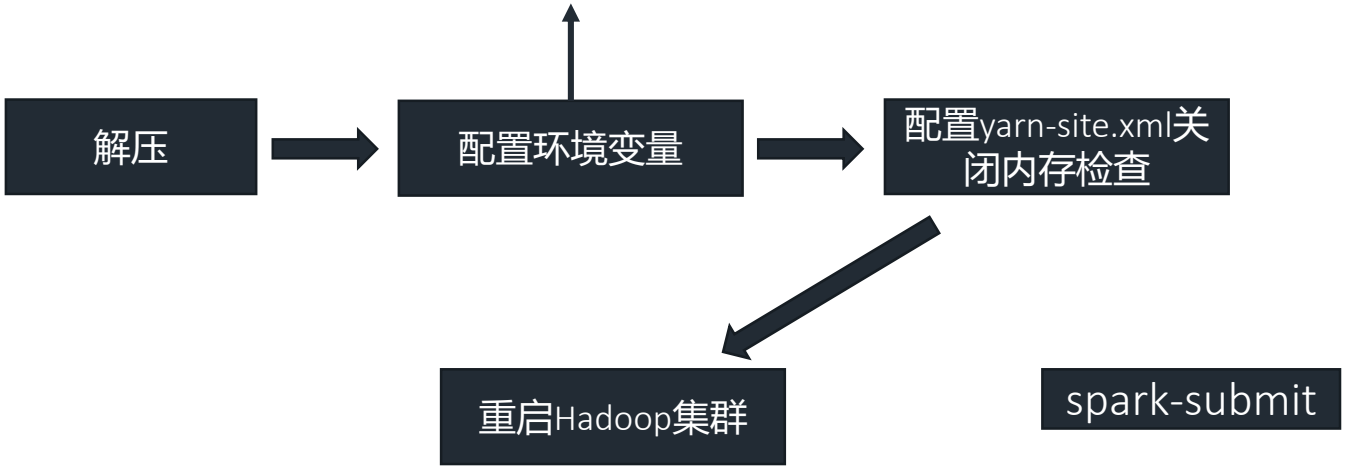
core-site.xml(**指定**namenode**地址**;**使用**hadoop**时产生文件的存放目录等)；** hdfs-site.xml(**副本数量、**hdfs**权限设置等)；** yarn-site.xml(yarn**地址、对外暴露的接口等)；配置节点目录文件；**3)**将**Hadoop**解压包发送给节点；**

搭建Spark集群

- On Yarn模式
- 如何运行

On Yarn模式搭建

```
export HADOOP_CONF_DIR=$HADOOP_HOME/etc/hadoop
export HDFS_CONF_DIR=$HADOOP_HOME/etc/hadoop
export YARN_CONF_DIR=$HADOOP_HOME/etc/hadoop
export PATH=$PATH:/root/spark-2.1.1-bin-hadoop2.7/bin
```



Spark Core程序编写

Scala语言 Spark RDD

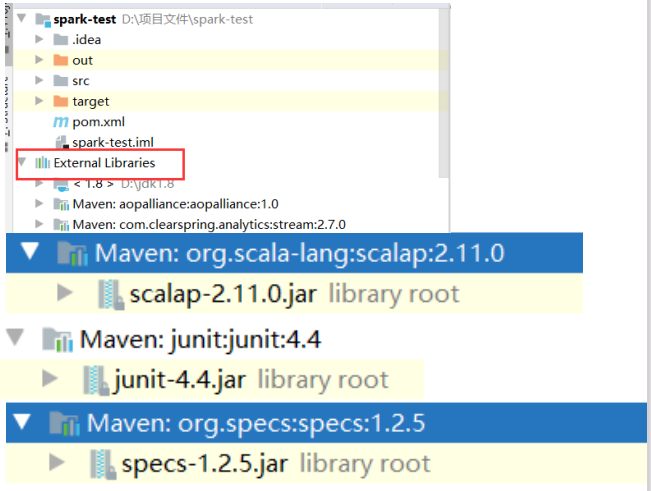
Spark Core程序编写

Scala是Spark常用的编程语言之一，是一门多范式编程语言。Scala被编译成Java字节码，所以它可以运行于JVM之上，并可以调用现有的java类库。

使用IDEA创建工程 → 编写程序 → 程序打包

Maven项目对象模型(POM)，可以通过一小段描述信息来管理项目的构建，报告和文档的项目管理工具软件。由于 Maven 的缺省构建规则有较高的可重用性，所以常常用两三行 Maven 构建脚本就可以构建简单的项目。

```
<dependencies>
  <dependency>
    <groupId>org.scala-lang</groupId>
    <artifactId>scala-library</artifactId>
    <version>${scala.version}</version>
  </dependency>
  <dependency>
    <groupId>junit</groupId>
    <artifactId>junit</artifactId>
    <version>4.4</version>
    <scope>test</scope>
  </dependency>
  <dependency>
    <groupId>org.specs</groupId>
    <artifactId>specs</artifactId>
    <version>1.2.5</version>
    <scope>test</scope>
  </dependency>
```

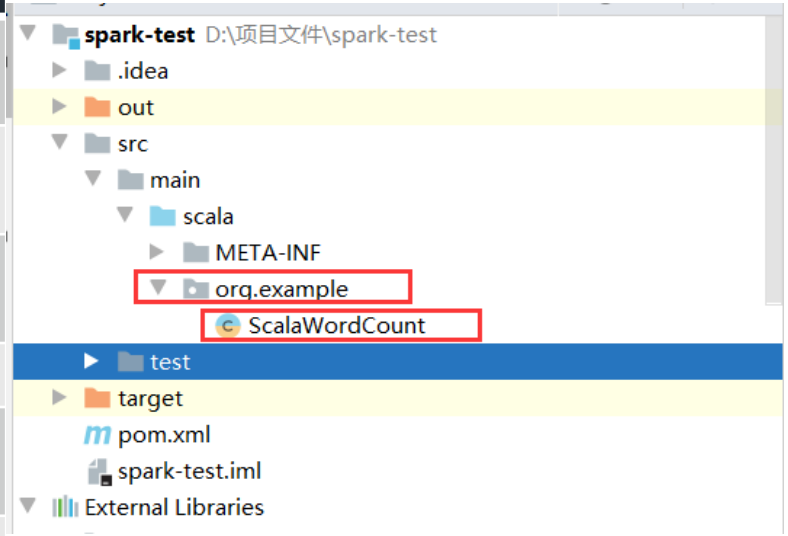


程序运行

程序运行

spark-submit --class org.example.ScalaWordCount --master yarn --num-executors 3 --driver-memory 1g --executor-memory 1g --executor-cores 1 spark-test.jar

参数	含义
--class	应用程序的主类，仅针对 java 或 scala 应用；
--master	master 的地址，提交任务到哪里执行，例如 spark://host:port，yarn，local；
--num-executors	启动的 executor 数量。默认为2。在 yarn 下使用；
--driver-memory	Driver内存，默认 1G；
--executor-memory	每个 executor 的内存，默认是1G；
--executor-core	每个 executor 的核数。在yarn或者standalone下使用；



运行结果

```
21/04/14 10:46:04 INFO scheduler.DAGScheduler: Job 1 finished: collect at ScalaWordCount.scala:19, took 0.080027
(hi,6),(hello,5),(spark,2),(sparkkgraphx,1),(sparkstreaming,1),(sparksql,1)21/04/14 10:46:04 INFO storage.BlockMa
t 3 piece0 on 192.168.0.138:44381 in memory (size: 2023.0 B, free: 366.3 MB)
```

```
[root@master ~]# hadoop fs -ls /
21/04/12 11:51:17 WARN util.NativeCodeLoader: Unable to load native-ha
re applicable
Found 3 items
drwxr-xr-x   - root supergroup          0 2021-04-12 11:49 /spark_test
drwx------   - root supergroup          0 2021-04-11 22:49 /tmp
drwxr-xr-x   - root supergroup          0 2021-04-11 22:49 /user
[root@master ~]#
```

```
[root@master ~]# hadoop fs -cat /spark_test/part-00000
21/04/12 11:51:39 WARN util.NativeCodeLoader: Unable to load native-hadoop lib
re applicable
(hi,6)
(hello,5)
(spark,2)
[root@master ~]# ifconfig
eth0: flags=4163<UP,BROADCAST,RUNNING,MULTICAST>  mtu 1500
        inet 192.168.0.138  netmask 255.255.255.0  broadcast 192.168.0.255
        inet6 fe80::f816:3eff:fe28:412  prefixlen 64  scopeid 0x20<link>
```

hadoop

Logged in as: dr.who

FINISHED Applications

- Cluster
  - About
  - Nodes
  - Node Labels
  - Applications
    - NEW
    - NEW_SAVING
    - SUBMITTED
    - ACCEPTED
    - RUNNING
    - FINISHED
    - FAILED
    - KILLED
  - Scheduler
- Tools

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	VCores Used	VCores Total	VCores Reserved	Active Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes
4	0	0	4	0	0 B	16 GB	0 B	0	16	0	2	0	0	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation
Capacity Scheduler	[MEMORY]	<memory:1024, vCores:1>	<memory:8192, vCores:8>

Show 20 entries Search:

ID	User	Name	Application Type	Queue	StartTime	FinishTime	State	FinalStatus	Progress	Tracking UI	Blacklisted Nodes
application_1618366973845_0004	root	word-count	SPARK	default	Wed Apr 14 11:55:54 +0800 2021	Wed Apr 14 11:56:09 +0800 2021	FINISHED	SUCCEEDED		History	N/A
application_1618366973845_0003	root	word-count	SPARK	default	Wed Apr 14 11:53:35 +0800 2021	Wed Apr 14 11:53:52 +0800 2021	FINISHED	SUCCEEDED		History	N/A
application_1618366973845_0002	root	Spark Pi	SPARK	default	Wed Apr 14 11:04:39 +0800 2021	Wed Apr 14 11:04:52 +0800 2021	FINISHED	SUCCEEDED		History	N/A
application_1618366973845_0001	root	Spark Pi	SPARK	default	Wed Apr 14 10:37:37 +0800 2021	Wed Apr 14 10:37:53 +0800 2021	FINISHED	SUCCEEDED		History	N/A

Showing 1 to 4 of 4 entries First Previous 1 Next Last

注意事项

注意事项

序号	问题	解决方案
1	关于环境	在客户端打包程序的时候，使用实验提供的环境，要求jdk版本为1.8，否则会出错；
2	新建类，无Scala.class	https://blog.csdn.net/qq_16410733/article/details/85038832
3	关于防火墙	一定要检查三个节点是否关闭了防火墙；
4	文件拷贝	注意主节点配置的文件是否全部拷贝给了数据节点，主要包括(Hadoop文件夹、hosts、.bash_profile)
5	Hadoop需要进一步验证是否能够正常使用	运行：hadoop jar ./hadoop-2.7.3/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.7.3.jar pi 10 10 输出结果：3.2000
6	关于web页面	web页面默认在本地端是无法打开的，需要配置华为云的安全组，开放指定的端口，如：50070

感谢聆听