4.Nutch安装

安装环境：centos 6.5
nutch：v2.2.1
hbase:v0.94.18
本篇重点讲述nutch的安装和nutch与hbase的集成，hbase的安装请参考其他资料；
安装步骤：

1. 安装ant：因编译nutch源码，需要ant工具，下载apache-ant 设置 系统变量
写道
[hadoop@master nutch]$ vim /etc/profile
添加：ANT_HOME=/usr/local/ant 变量，并将AN_HOMT添加到PATH

2. 下载nutch安装包：http://nutch.apache.org/downloads.html，下载目前最新的apache-nutch-2.2.1-src.tar.gz

```
apache-nutch-2.2.1-src.tar.gz
```

```
[hadoop@master local]$ chmod 777 apache-nutch-2.2.1-src.tar.gz
[hadoop@master local]$ tar zxvf apache-nutch-2.2.1-src.tar.gz
[hadoop@master local]$ mv apache-nutch-2.2.1 nutch
[hadoop@master local]$ cd nutch/
```

3. 修改nutch的conf/nutch-site.xml文件，添加如下代码：

```
<property>
     <name>storage.data.store.class</name>
     <value>org.apache.gora.hbase.store.HBaseStore</value>
     <description>Default class for storing data</description>
</property>
<property>
     <name>http.agent.name</name>
     <value>Mozilla/5.0 (Macintosh; Intel Mac OS X 10_8_4) AppleWebKit/537.36 (KHTML, like Gecko)
     Chrome/28.0.1500.95 Safari/537.36</value>
</property>
```

4. 修改ivy/ivy.xml文件，找到：
开启`<dependency org="org.apache.gora" name="gora-hbase" rev="0.3" conf="*->default" />`

5. 修改conf/gora.properies：

```
gora.datastore.default=org.apache.gora.hbase.store.HBaseStore
#gora.sqlstore.jdbc.driver=org.hsqldb.jdbc.JDBCDriver
#gora.sqlstore.jdbc.url=jdbc:hsqldb:hsql://localhost/nutchtest
#gora.sqlstore.jdbc.user=sa
#gora.sqlstore.jdbc.password=
```

6. ant编译nutch：切换到nutch目录：

```
[hadoop@master local]$ cd nutch
[hadoop@master nutch]$ ant
```

7. 编译过程会等待一段时间。
修改nutch配置文件：在编译nutch源文件前，为了支持hbase存储，需要修改相应的配置：

```
替换hbase-site.xml 和 jar
路径如下：nutch/runtime/conf  nutch/runtime/lib
    #拷贝hbase的配置文件到nutch
    cp /usr/local/hbase/conf/hbase-site.xml /usr/local/nutch/runtime/conf/
```

1. 复制hbase的jar包到nutch，本人安装的hbase是hbase0.94.18，nutch自带的gora0.3是只能支持到最高hbase0.92，默认是hbase0.90，而默认的0.90jar包去操作0.94的hbase，导致一个异常：
2. Java代码

```
java.lang.IllegalArgumentException: Not a host:port pair
```

1. 应该是低版本hbase client操作高版本hbase server的常见错误，但也不能直接用0.94的hbase jar包去替换，不然又会导致另一个错误：
2. java.lang.NoSuchMethodError:org.apache.hadoop.hbase.HColumnDescriptor.setMaxVersions(I)V

1. 解决办法：我们选择hbase 0.92 到 0.93之间的版本，首先尝试0.92版本，可以从maven中心库下载：
2. http://central.maven.org/maven2/org/apache/hbase/hbase/0.92.2/hbase-0.92.2.jar

然后将hbase-0.92.2.jar包替换nutch/runtime/lib

设置抓取网址： 编译后切换到目录：

1. Java代码

```
[hadoop@master nutch]$ cd runtime/local/
[hadoop@master local]$ mkdir -p urls
```

```
[hadoop@master local]$ vim urls/seed.txt
```

1. 填写seed.txt内容： http://www.apache.org/ 每一行为一个目标地址；并将urls目录放到hdfs文件系统上：
2. Java代码

```
hadoop fs -copyFromLocal urls /home/hadoop/urls
```

运行nutch测试：执行nutch inject将网页种子放到hbase中

1. Java代码

```
[hadoop@master local]$ bin/nutch inject ./urls(本地路径)
```

1. 查看hbase中表：
2. Java代码

```
hbase shell
```

进入到hbaseshell后查看表
>list

1. 看到有表“webpage”则表示成功；

然后一次执行

1. Java代码

```
[hadoop@master local]$ bin/nutch  generate -tpN 3
[hadoop@master local]$ bin/nutch fetch -all
[hadoop@master local]$ bin/nutch parse -all
[hadoop@master local]$ bin/nutch updatedb
```

切换到hbase shell或使用hbase client查看数据
bin/nutch crawl urls -depth 3 -topN 5 -佳佳给的命令