

HDFS应用实践

——WebHDFS



联创中控（北京）科技有限公司

本节要点

大纲

1 / WebHDFS的配置

2 / WebHDFS命令



WebHDFS的原理

使用curl命令向指定的Hadoop集群对外接口发送页面请求，Hadoop集群的网络接口接收到请求之后，会将命令中数据解析成HDFS上对应文件或者文件夹、用户、权限、缓存大小等参数。待完成相应的操作之后，将结果发还给执行curl命令的客户端，并显示执行信息或者错误信息。



WebHDFS的配置方式

修改Hadoop集群的hdfs-site.xml配置，使其开放WebHDFS服务

修改hdfs-site.xml配置为：

```
<property>  
<name>dfs.webhdfs.enabled</name>  
<value>true</value>  
</property>  
<property>  
<name>dfs.permission</name>  
<value>>false</value>  
</property>
```

本节要点

大纲

1 / WebHDFS的配置

2 / WebHDFS命令



WebHDFS的语法形式

```
curl [-i/-X/-u/-T] [PUT] "HTTP://<HOST>:<PORT>/webhdfs/v1  
/<PATH>?[user.name=<user>&]&op=<operation>&[doas=<u  
ser>] ..."
```

- http : //<HOST>:<PORT>代表需要将命令发送的地址和端口，也就是Hadoop集群服务器的IP地址和HDFS端口（默认是50070）。
- 在这个地址之后的部分/webhdfs/v1/<PATH>代表着需要操作的远程HDFS集群上的路径。
- op参数，代表着具体的操作指令。



WebHDFS命令：创建文件并写入内容

```
curl -i -X PUT "http://<HOST>:<PORT>/webhdfs/v1/<PATH>?op=CREATE [overwrite=<true | false>] [&blocksize=<LONG>][&replication=<SHORT>][&permission=<OCTAL>] [&buffersize=<INT>]"
```

返回值
location包
括了已创
建文件所
在的
DataNode
地址及创
建路径。

```
[root@master1 ~]# curl -i -X PUT "http://master1:50070/webhdfs/v1/webhdfsFile?op=CREATE"
HTTP/1.1 307 TEMPORARY_REDIRECT
Cache-Control: no-cache
Expires: Thu, 05 Nov 2015 08:13:35 GMT
Date: Thu, 05 Nov 2015 08:13:35 GMT
Pragma: no-cache
Expires: Thu, 05 Nov 2015 08:13:35 GMT
Date: Thu, 05 Nov 2015 08:13:35 GMT
Pragma: no-cache
Content-Type: application/octet-stream
Location: http://master1:50075/webhdfs/v1/webhdfsFile?op=CREATE&namenoderpcaddress=master1:8020&overwrite=false
Content-Length: 0
Server: Jetty(6.1.26)
```



WebHDFS命令：创建文件并写入内容

结合返回的地址发送数据

```
curl -i -X PUT -T <LOCAL_FILE> " http:<DATANODE>:<PORT>/webhdfs/  
v1/<PATH>?op=CREATE..."
```

```
[root@master1 ~]# curl -i -X PUT -T /usr/mytext/1.txt "http://master1:50075/web  
hdfs/v1/webhdfsFile?op=CREATE&namenoderpcaddress=master1:8020&overwrite=false"  
HTTP/1.1 100 Continue  
  
HTTP/1.1 201 Created  
Location: hdfs://master1:8020/webhdfsFile  
Content-Length: 0  
Connection: close
```



WebHDFS命令：向文件追加内容

使用下面的命令获取待追加内容文件所在地址

```
curl -i -X POST " http://<HOST>:<PORT>/webhdfs/v1/<PATH>?op=APPEND[&bufferize=<INT>]"
```

```
[root@master1 ~]# curl -i -X POST "http://master1:50070/webhdfs/v1/webhdfsFile?op=APPEND"
HTTP/1.1 307 TEMPORARY_REDIRECT
Cache-Control: no-cache
Expires: Thu, 05 Nov 2015 08:28:24 GMT
Date: Thu, 05 Nov 2015 08:28:24 GMT
Pragma: no-cache
Expires: Thu, 05 Nov 2015 08:28:24 GMT
Date: Thu, 05 Nov 2015 08:28:24 GMT
Pragma: no-cache
Content-Type: application/octet-stream
Location: http://master1:50075/webhdfs/v1/webhdfsFile?op=APPEND&namenoderpcaddress=master1:8020
Content-Length: 0
Server: Jetty(6.1.26)
```



WebHDFS命令：向文件追加内容

结合返回内容的location信息，追加内容，命令如下

```
curl -i -X POST -T <LOCAL_FILE> " http:<DATANODE>:<PORT>/webhdfs/v1/  
<PATH>?op=APPEND"
```

```
[root@master1 ~]# curl -i -X POST -T /usr/mytext/2.txt "http://master1:50075/we  
bhdfs/v1/webhdfsFile?op=APPEND&namenoderpcaddress=master1:8020"  
HTTP/1.1 100 Continue  
  
HTTP/1.1 200 OK  
Content-Length: 0  
Connection: close
```



WebHDFS命令：打开并读取内容

打开并读取文件内容，使用下面的命令打开远程HDFS上的文件并读取内容：

```
curl -i -L " http://<HOST>:<PORT>/webhdfs/v1/<PATH>?op=
OPEN[&offset=<LONG>][&bufferize=<INT>]"
```



WebHDFS命令：创建文件夹

创建文件夹：

```
curl -i -X PUT " http://<HOST>:<PORT>/webhdfs/v1/<PATH>?op=MKDIRS[&permission=<OCTAL>]"
```

```
[root@master1 ~]# curl -i -X PUT "http://master1:50070/webhdfs/v1/webhdfsFile2?op=MKDIRS"
HTTP/1.1 200 OK
Cache-Control: no-cache
Expires: Thu, 05 Nov 2015 08:51:13 GMT
Date: Thu, 05 Nov 2015 08:51:13 GMT
Pragma: no-cache
Expires: Thu, 05 Nov 2015 08:51:13 GMT
Date: Thu, 05 Nov 2015 08:51:13 GMT
Pragma: no-cache
Content-Type: application/json
Transfer-Encoding: chunked
Server: Jetty(6.1.26)

{"boolean":true}[root@master1 ~]#
[root@master1 ~]# hadoop fs -ls /
15/11/05 16:51:29 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 4 items
drwxr-xr-x  - root  supergroup          0 2015-11-02 15:24 /master1-file
-rw-r--r--  1 root  supergroup    107902 2015-11-03 11:20 /master1-file2
-rwxr-xr-x  1 dr.who supergroup    215817 2015-11-05 16:37 /webhdfsFile
drwxr-xr-x  - dr.who supergroup          0 2015-11-05 16:51 /webhdfsFile2
[root@master1 ~]#
```



WebHDFS命令：重命名文件或文件夹

重命名文件夹或文件：

```
curl -i -X PUT "http://<HOST>:<PORT>/webhdfs/v1/<PATH>?op=
RENAME&destination=<PATH>"
```

```
[root@master1 ~]# curl -i -X PUT "http://master1:50070/webhdfs/v1/webhdfsFile2?op=
p=RENAME&destination=/webhdfsFile3"
HTTP/1.1 200 OK
Cache-Control: no-cache
Expires: Thu, 05 Nov 2015 08:54:24 GMT
Date: Thu, 05 Nov 2015 08:54:24 GMT
Pragma: no-cache
Expires: Thu, 05 Nov 2015 08:54:24 GMT
Date: Thu, 05 Nov 2015 08:54:24 GMT
Pragma: no-cache
Content-Type: application/json
Transfer-Encoding: chunked
Server: Jetty(6.1.26)

{"boolean": true}[r^C
[root@master1 ~]# hadoop fs -ls /
15/11/05 16:54:43 WARN util.NativeCodeLoader: Unable to load native-hadoop library
for your platform... using builtin-java classes where applicable
Found 4 items
drwxr-xr-x   - root    supergroup          0 2015-11-02 15:24 /master1-file
-rw-r--r--   1 root    supergroup       107902 2015-11-03 11:20 /master1-file2
-rwxr-xr-x   1 dr.who  supergroup       215817 2015-11-05 16:37 /webhdfsFile
drwxr-xr-x   - dr.who  supergroup          0 2015-11-05 16:51 /webhdfsFile3
[root@master1 ~]#
```



WebHDFS命令：删除文件或文件夹

删除文件夹或者文件：

```
curl -i -X DELETE "http://<host>:<port>/webhdfs/v1/<PATH>?op=DELETE [&recursive=<true | false>]"
```

```
[root@master1 ~]# curl -i -X DELETE "http://master1:50070/webhdfs/v1/webhdfsFile3?op=DELETE"
HTTP/1.1 200 OK
Cache-Control: no-cache
Expires: Thu, 05 Nov 2015 08:58:07 GMT
Date: Thu, 05 Nov 2015 08:58:07 GMT
Pragma: no-cache
Expires: Thu, 05 Nov 2015 08:58:07 GMT
Date: Thu, 05 Nov 2015 08:58:07 GMT
Pragma: no-cache
Content-Type: application/json
Transfer-Encoding: chunked
Server: Jetty(6.1.26)

[ root@master1 ~]# hadoop fs -ls /
15/11/05 16:59:59 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 3 items
drwxr-xr-x  - root    supergroup          0 2015-11-02 15:24 /master1-file
-rw-r--r--  1 root    supergroup    107902 2015-11-03 11:20 /master1-file2
-rwxr-xr-x  1 dr.who  supergroup    215817 2015-11-05 16:37 /webhdfsFile
```



WebHDFS命令：查看文件或文件夹信息

查看文件夹或文件信息：

```
curl -i "http://<HOST>:<PORT>/webhdfs/v1/<PATH>?op=GETFILESTATUS"
```

```
[root@master1 ~]# curl -i "http://master1:50070/webhdfs/v1/webhdfsFile?op=GETFILESTATUS"
HTTP/1.1 200 OK
Cache-Control: no-cache
Expires: Thu, 05 Nov 2015 09:03:00 GMT
Date: Thu, 05 Nov 2015 09:03:00 GMT
Pragma: no-cache
Expires: Thu, 05 Nov 2015 09:03:00 GMT
Date: Thu, 05 Nov 2015 09:03:00 GMT
Pragma: no-cache
Content-Type: application/json
Transfer-Encoding: chunked
Server: Jetty(6.1.26)

{"FileStatus": {"accessTime": 1446711540079, "blockSize": 134217728, "childrenNum": 0, "fileId": 16409, "group": "supergroup", "length": 215817, "modificationTime": 1446712673171, "owner": "dr. who", "pathSuffix": "", "permission": "755", "replication": 1, "storagePolicy": 0, "type": "FILE"}}[root@master1 ~]#
```



WebHDFS命令：列举文件夹内容

列举文件夹内容：

```
curl -i "http://<HOST>:<PORT>/webhdfs/v1/<PATH>?op=LISTSTATUS"
```

```
[root@master1 ~]# curl -i "http://master1:50070/webhdfs/v1/webhdfsFile?op=LISTSTATUS"
HTTP/1.1 200 OK
Cache-Control: no-cache
Expires: Thu, 05 Nov 2015 09:07:34 GMT
Date: Thu, 05 Nov 2015 09:07:34 GMT
Pragma: no-cache
Expires: Thu, 05 Nov 2015 09:07:34 GMT
Date: Thu, 05 Nov 2015 09:07:34 GMT
Pragma: no-cache
Content-Type: application/json
Transfer-Encoding: chunked
Server: Jetty(6.1.26)

{"FileStatuses":{"FileStatus":[{"accessTime":1446711540079,"blockSize":134217728,"childrenNum":0,"fileId":16409,"group":"supergroup","length":215817,"modificationTime":1446712678171,"owner":"dr.who","pathSuffix":"","permission":"755","replication":1,"storagePolicy":0,"type":"FILE"}]}
```



WebHDFS命令：获取文件夹统计信息

获取文件夹统计信息：

```
curl -i "http://<HOST>:<PORT>/webhdfs/v1/<PATH>?op=GETCONTENTSUMMARY"
```

```
[root@master1 ~]# curl -i "http://master1:50070/webhdfs/v1/webhdfsFile?op=GETCONTENTSUMMARY"
HTTP/1.1 200 OK
Cache-Control: no-cache
Expires: Thu, 05 Nov 2015 09:04:30 GMT
Date: Thu, 05 Nov 2015 09:04:30 GMT
Pragma: no-cache
Expires: Thu, 05 Nov 2015 09:04:30 GMT
Date: Thu, 05 Nov 2015 09:04:30 GMT
Pragma: no-cache
Content-Type: application/json
Transfer-Encoding: chunked
Server: Jetty(6.1.26)

{"ContentSummary": {"directoryCount": 0, "fileCount": 1, "length": 215817, "quota": -1, "spaceConsumed": 215817, "spaceQuota": -1}}[root@master1 ~]#
```

WebHDFS命令：获取文件校验和

获取文件校验和：

```
curl -i "http://<HOST>:<PORT>/webhdfs/v1/<PATH>?op=GETFILECHECKSUM"
```

```
[root@master1 ~]# curl -i "http://master1:50075/webhdfs/v1/webhdfsFile?op=GETFILECHECKSUM&namenoderpcaddress=master1:8020"
HTTP/1.1 200 OK
Content-Type: application/json; charset=utf-8
Content-Length: 144
Connection: close

{"FileChecksum":{"algorithm":"MD5-of-0MD5-of-512CRC32C","bytes":"00000200000000000000000009bd886c9dc574d8edcc39d69695be62400000000","length":28}}[root@master1 ~]#
```



WebHDFS命令：获取当前web用户主目录

获取当前web用户的主目录：

```
curl -i "http://<HOST>:<PORT>/webhdfs/v1/?op=GETHOMEDIRECTORY"
```

```
[root@master1 ~]# curl -i "http://master1:50070/webhdfs/v1/?op=GETHOMEDIRECTORY"
HTTP/1.1 200 OK
Cache-Control: no-cache
Expires: Thu, 05 Nov 2015 09:13:27 GMT
Date: Thu, 05 Nov 2015 09:13:27 GMT
Pragma: no-cache
Expires: Thu, 05 Nov 2015 09:13:27 GMT
Date: Thu, 05 Nov 2015 09:13:27 GMT
Pragma: no-cache
Content-Type: application/json
Transfer-Encoding: chunked
Server: Jetty(6.1.26)

{"Path":"/user/dr.who"}[root@master1 ~]#
```



WebHDFS命令：设置文件操作权限

设置权限：

```
curl -i -X PUT " http://<HOST>:<PORT>/webhdfs/v1/<PATH>?op=
SETPERMISSION [&permission=<OCTAL>]"
```

```
[root@master1 ~]# curl -i -X PUT "http://master1:50070/webhdfs/v1/webhdfsFile?op=
=SETPERMISSION"
HTTP/1.1 200 OK
Cache-Control: no-cache
Expires: Thu, 05 Nov 2015 09:16:50 GMT
Date: Thu, 05 Nov 2015 09:16:50 GMT
Pragma: no-cache
Expires: Thu, 05 Nov 2015 09:16:50 GMT
Date: Thu, 05 Nov 2015 09:16:50 GMT
Pragma: no-cache
Content-Type: application/octet-stream
Content-Length: 0
Server: Jetty(6.1.26)
```



WebHDFS命令：设置备份数量

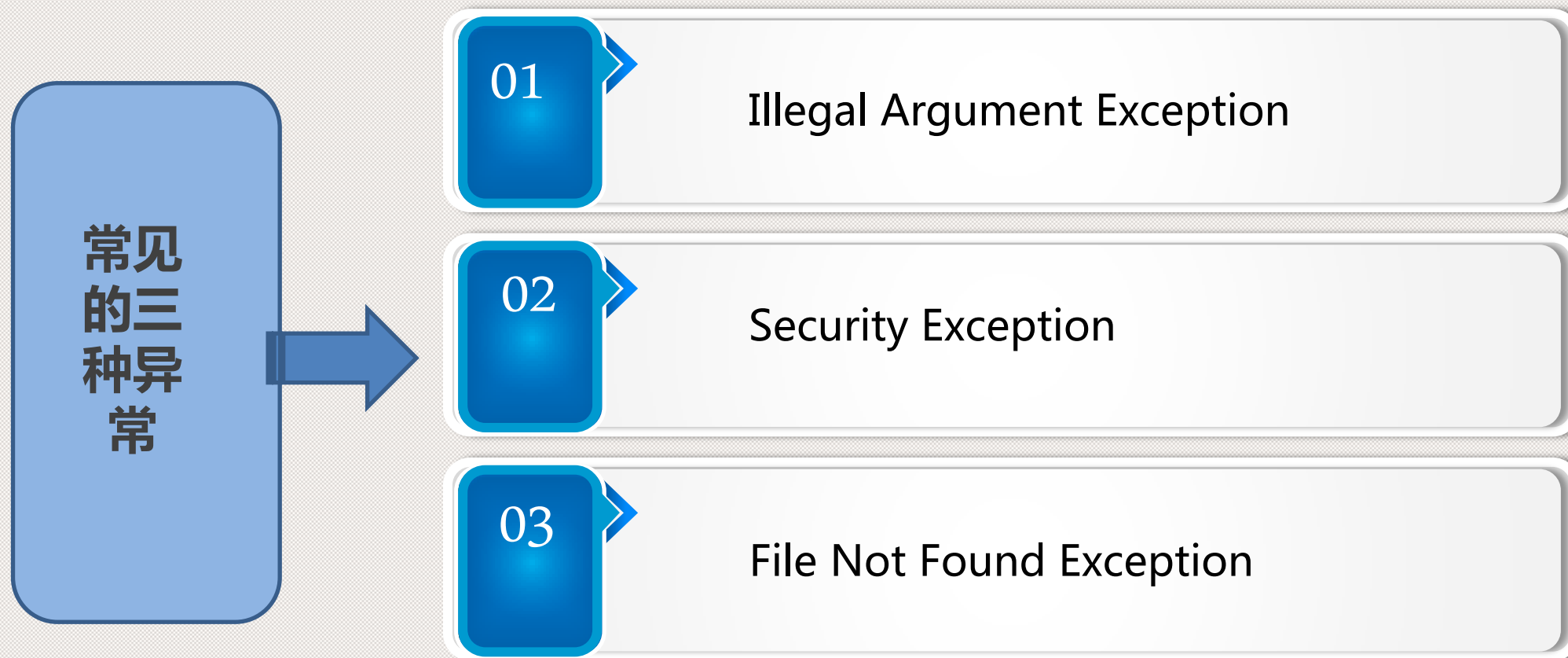
设置备份数量：

```
curl -i -X PUT "http://<HOST>:<PORT>/webhdfs/v1/<PATH>?op=SETREPLICATION [&replication=<SHORT>]"
```

```
[root@master1 ~]# curl -i -X PUT "http://master1:50070/webhdfs/v1/webhdfsFile?op=SETREPLICATION"
HTTP/1.1 200 OK
Cache-Control: no-cache
Expires: Thu, 05 Nov 2015 09:23:01 GMT
Date: Thu, 05 Nov 2015 09:23:01 GMT
Pragma: no-cache
Expires: Thu, 05 Nov 2015 09:23:01 GMT
Date: Thu, 05 Nov 2015 09:23:01 GMT
Pragma: no-cache
Content-Type: application/json
Transfer-Encoding: chunked
Server: Jetty(6.1.26)
```



常见异常分析



常见异常分析

Illegal Argument Exception

```
HTTP/1.1 400 Bad Request
Transfer Encoding: chunked
{
  "RemoteException":
  {
    "exception" : "IllegalArgumentException",
    "javaClassName": "java.lang.IllegalArgumentException",
    "message": "Invalid value for webhdfs parameter \permission\"
  }
}
```

从异常信息可以很明显看出是命令的参数不对，这就需要仔细检查自己的参数是否有输入错误或拼写错误。



Security Exception

```
HTTP/1.1 400 Unauthorized
Content-Type: application/json
Transfer-Encoding: chunked
{
  "RemoteException":
  {
    "exception" : "SecurityException",
    "javaClassName": "java.lang. SecurityException",
    "message": "Failed to obtain user group information"
  }
}
```

出现这种异常的原因是提交命令的用户应通过认证，这就需要提交认证信息。

常见异常分析

File Not Found Exception

```
HTTP/1.1 404 Not Found
Content-Type: application/json
Transfer-Encoding: chunked
{
  "RemoteException":
  {
    "exception" : "FileNotFoundException",
    "javaClassName": "java.io. FileNotFoundException",
    "message": "File does not exist: /a"
  }
}
```

出现这种异常的原因就是找不到指定的目录或者文件，这需要确认自己命令的路径和文件。



THANKS

谢 谢 观 看



联创中控（北京）科技有限公司

地址：北京市昌平区北京国际信息产业基地高新四街6号院5层

电话：400-659-9866

网址：<http://www.uicctech.com>