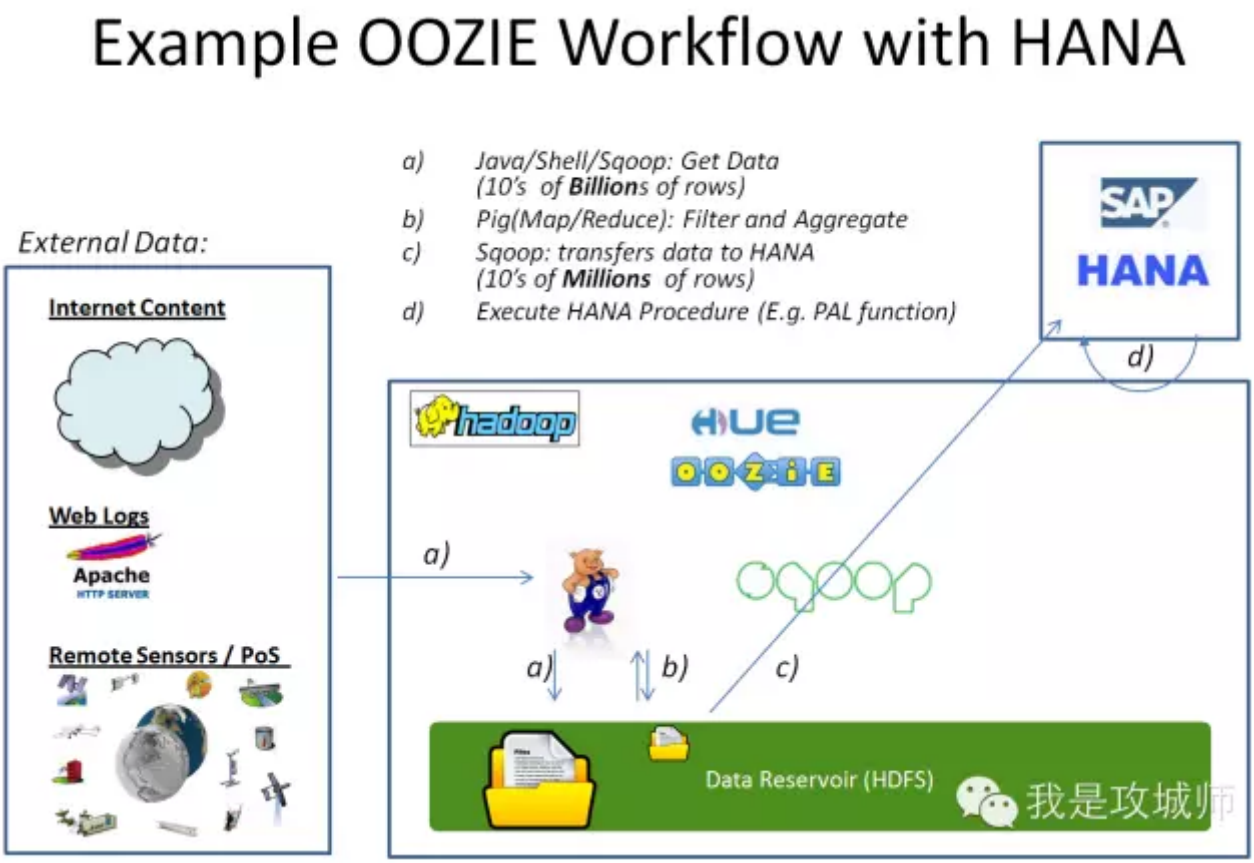
驭象者之Apache Oozie

2015-05-19 我是攻城师

（1）Apache Oozie是什么？

Oozie在英语中的释义指的是：驯象人，驭象者（多指缅甸那边的俗称），这个比喻相对与它的功能来说，还是很恰当的。

Apache Oozie是一个用来管理Hadoop任务的工作流调度系统，是基于有向无环图的模型（DAG）。Oozie支持大多数的Hadoop任务的组合，常见的有Java MapReduce，Streaming map-reduce，Pig，Hive， Sqoop ，Distcp，也可以结合一些脚本如Shell，Python，Java来很灵活的完成一些事情。同时，它也是一个可伸缩的，可扩展，高可靠的的系统



（2）Apache Oozie能用来干什么？

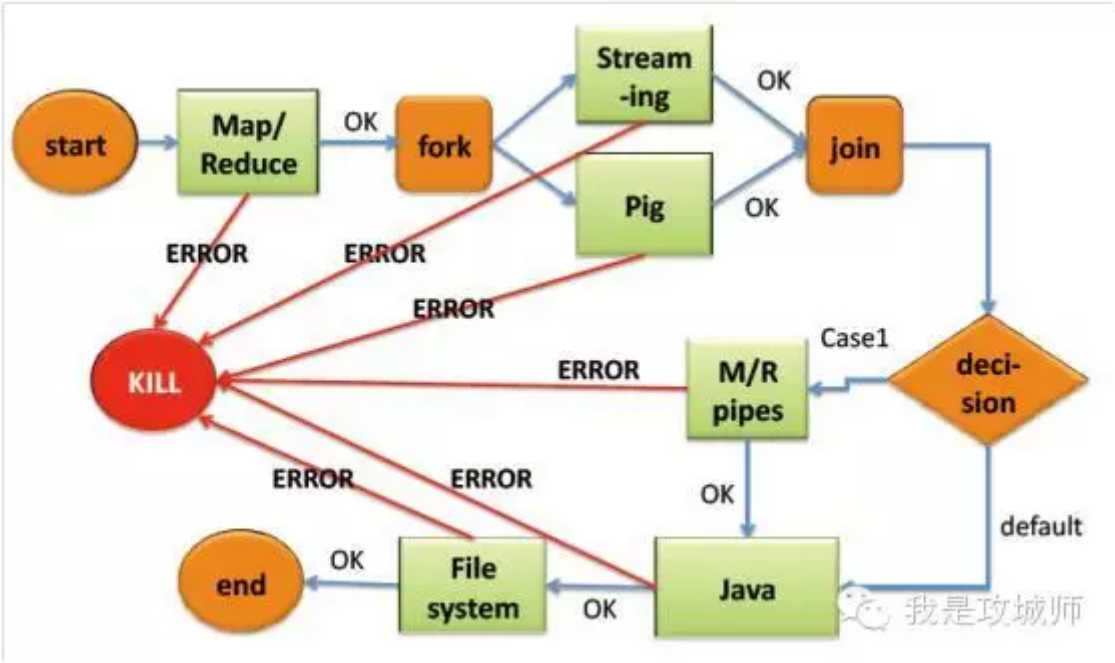
其实，上面的这张图，已经足够回答这个问题了，工作流嘛，顾名思义，就是我要干一件事，需要很多步骤，然后有序组合，最终达到能够完成这件事的目的。

举个例子，就拿做饭这件事吧。

1，买菜
2，洗菜
3，切菜
4，炒菜
5，上菜

这是一个简单的流程，当然这里面会有很多其他的小细节，比如我买菜，去了不同的菜市场，炒菜时候，又临时去买了一些调料，等等。

仔细分析这里面的道道，有些是有依赖关系的，有些没依赖关系的，比如菜是核心，所有很菜有关的都有先后顺序，其他的辅助步骤，比如说烧水，跟这是没有依赖关系的。反应到实际工作中的一些任务也是如此，所以采用oozie来管理调度，还是很方便的一件事。



（3）Oozie的组成

Readme, license, notice & Release log files.（一个项目的，版权，介绍，log等）

Oozie server: oozie-server directory.（oozie的服务端目录）

Scripts: bin/ directory, client and server scripts.（bin下面有一些常用的命令，来管理oozie的）

Binaries: lib/ directory, client JAR files.（存放oozie的依赖包）

Configuration: conf/ server configuration directory.（oozie的配置文件）

Archives:（归档包目录）

oozie-client-*.tar.gz : Client tools.（oozie的客户端包）

oozie.war : Oozie WAR file.（web的服务工程）

docs.zip : Documentation.（文档）

oozie-examples-*.tar.gz : Examples.（例子）

oozie-sharelib-*.tar.gz : Share libraries (with Streaming, Pig JARs).(一些工作流支持的框架共享包)

（4）oozie支持调度的应用

1，Email任务

2，Shell任务

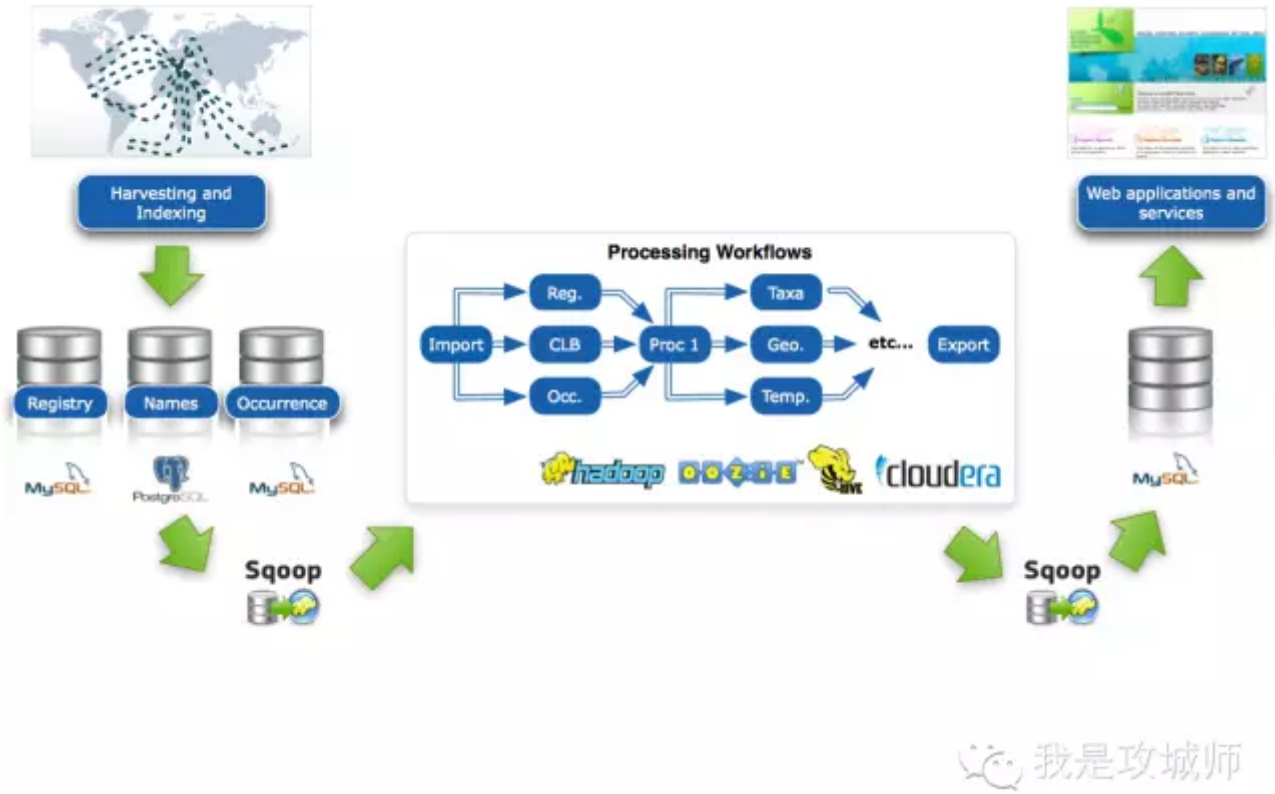
3，Hive任务

4，Sqoop任务

5，SSH任务

6，Distcp任务

7，自定义的任务



（5）oozie的下载，安装，编译

oozie目前最新的版本是oozie4.1.0，下载地址1，如果链接不上，可点击这个下载地址2，

在linux上，可以直接wget http://archive.apache.org/dist/oozie/4.1.0/oozie-4.1.0.tar.gz下载

下载完，可以解压出来根据自己的一些环境编译。

散仙这里的一些环境如下：

Hadoop2.2
JDK1.7
Maven3.0.5
Ant1.9.4
Hive0.13.1
Pig0.12.1

所以，需要修改在oozie的根目录下的pom文件：

1，修改JDK版本

2，如有必要可修改各个组件的版本，在跟目录下执行

grep -l "2.3.0" `find . -name "pom.xml"`

Java代码

```
./pom.xml
./hadooplibs/hadoop-distcp-2/pom.xml
./hadooplibs/hadoop-test-2/pom.xml
./hadooplibs/hadoop-utils-2/pom.xml
./hadooplibs/hadoop-2/pom.xml
```

将查出来的pom文件，修改对应hadoop版本，hive，hbase，pig等组件版本

注意使用(sed -e 's/2.3.0/2.2.0/g' pom.xml 替换可能更快，但是建议自己去修改，因为改的地方并不是太多!)

注意，在4.1.0里，需要把下面这个保持成2.3.0，hadoop的版本可以是2.2.0如果，不改的话，编译Zookeeper-Scurity-Test时候，会报错

Java代码

```
[INFO] Apache Oozie ZooKeeper Security Tests ..... FAILURE [2.204s]
[INFO] ------------------------------------------------------------------------
[INFO] BUILD FAILURE
[INFO] ------------------------------------------------------------------------
[INFO] Total time: 5:27.818s
[INFO] Finished at: Fri May 15 12:50:50 CST 2015
[INFO] Final Memory: 132M/237M
[INFO] ------------------------------------------------------------------------
[ERROR] Failed to execute goal on project oozie-zookeeper-security-tests: Could not resolve dependencies for project org.apache.oozie:oozie-zookeeper-security-tests:jar:4.1.0: Failed to collect dependencies for [org.apache.curator:curator-test:jar:2.5.0 (test), org.apache.hadoop:hadoop-minikdc:jar:2.2.0 (test), org.apache.oozie:oozie-core:jar:4.1.0 (test), org.apache.oozie:oozie-core:jar:tests:4.1.0 (test), org.apache.oozie:oozie-hadoop:jar:2.2.0.oozie-4.1.0 (provided), org.apache.oozie:oozie-hadoop-test:jar:2.2.0.oozie-4.1.0 (test)]: Failed to read artifact descriptor for org.apache.hadoop:hadoop-minikdc:jar:2.2.0: Could not transfer artifact org.apache.hadoop:hadoop-minikdc:pom:2.2.0 from/to Codehaus repository (http://repository.codehaus.org/): peer not authenticated -> [Help 1]
[ERROR]
[ERROR] To see the full stack trace of the errors, re-run Maven with the -e switch.
[ERROR] Re-run Maven using the -X switch to enable full debug logging.
[ERROR]
[ERROR] For more information about the errors and possible solutions, please read the following articles:
[ERROR] [Help 1] http://cwiki.apache.org/confluence/display/MAVEN/DependencyResolutionException
[ERROR]
[ERROR] After correcting the problems, you can resume the build with the command
[ERROR] mvn <goals> -rf :oozie-zookeeper-security-tests
```

改回2.3.0即可

Java代码

```
1. <dependency>
2. <groupId>org.apache.hadoop</groupId>
3. <artifactId>hadoop-minikdc</artifactId>
4. <version>2.3.0</version>
5. </dependency>
```

3，修改完成后，执行下面命令进行编译：

bin/mkdistro.sh -DskipTests -Dhadoop.version=2.2.0

4，中间如果出现错误，不要紧，重新执行上面命令，会增量的编译，原来编译成功的，不会重复编译，编译成功如下：

Java代码

```
1. [INFO] Reactor Summary:
2. [INFO]
3. [INFO] Apache Oozie Main ..................................... SUCCESS [ 1.440 s]
4. [INFO] Apache Oozie Client ................................... SUCCESS [ 22.217 s]
5. [INFO] Apache Oozie Hadoop 1.1.1.oozie-4.1.0 ............... SUCCESS [ 0.836 s]
6. [INFO] Apache Oozie Hadoop Distcp 1.1.1.oozie-4.1.0 ....... SUCCESS [ 0.065 s]
7. [INFO] Apache Oozie Hadoop 1.1.1.oozie-4.1.0 Test ......... SUCCESS [ 0.182 s]
8. [INFO] Apache Oozie Hadoop Utils 1.1.1.oozie-4.1.0 ........ SUCCESS [ 0.784 s]
9. [INFO] Apache Oozie Hadoop 2.3.0.oozie-4.1.0 ............... SUCCESS [ 4.803 s]
10. [INFO] Apache Oozie Hadoop 2.3.0.oozie-4.1.0 Test ......... SUCCESS [ 0.254 s]
11. [INFO] Apache Oozie Hadoop Distcp 2.3.0.oozie-4.1.0 ....... SUCCESS [ 0.066 s]
12. [INFO] Apache Oozie Hadoop Utils 2.3.0.oozie-4.1.0 ........ SUCCESS [ 1.033 s]
13. [INFO] Apache Oozie Hadoop 0.23.5.oozie-4.1.0 ............ SUCCESS [ 3.231 s]
14. [INFO] Apache Oozie Hadoop 0.23.5.oozie-4.1.0 Test ........ SUCCESS [ 0.336 s]
15. [INFO] Apache Oozie Hadoop Distcp 0.23.5.oozie-4.1.0 ...... SUCCESS [ 0.062 s]
16. [INFO] Apache Oozie Hadoop Utils 0.23.5.oozie-4.1.0 ....... SUCCESS [ 0.878 s]
17. [INFO] Apache Oozie Hadoop Libs ........................... SUCCESS [ 3.780 s]
18. [INFO] Apache Oozie Hbase 0.94.2.oozie-4.1.0 ............... SUCCESS [ 0.338 s]
19. [INFO] Apache Oozie Hbase Libs ............................ SUCCESS [ 0.692 s]
20. [INFO] Apache Oozie HCatalog 0.13.1.oozie-4.1.0 ........... SUCCESS [ 0.919 s]
21. [INFO] Apache Oozie HCatalog Libs ......................... SUCCESS [ 1.735 s]
22. [INFO] Apache Oozie Share Lib Oozie ....................... SUCCESS [ 13.552 s]
23. [INFO] Apache Oozie Share Lib HCatalog .................... SUCCESS [ 40.232 s]
24. [INFO] Apache Oozie Core .................................. SUCCESS [05:03 min]
25. [INFO] Apache Oozie Docs .................................. SUCCESS [01:07 min]
```

```
[INFO] Apache Oozie Share Lib Pig ..... SUCCESS [01:38 min]
[INFO] Apache Oozie Share Lib Hive ..... SUCCESS [ 12.927 s]
[INFO] Apache Oozie Share Lib Sqoop ..... SUCCESS [ 5.655 s]
[INFO] Apache Oozie Share Lib Streaming ..... SUCCESS [ 4.577 s]
[INFO] Apache Oozie Share Lib Distcp ..... SUCCESS [ 1.900 s]
[INFO] Apache Oozie WebApp ..... SUCCESS [02:26 min]
[INFO] Apache Oozie Examples ..... SUCCESS [ 3.762 s]
[INFO] Apache Oozie Share Lib ..... SUCCESS [ 11.415 s]
[INFO] Apache Oozie Tools ..... SUCCESS [ 10.718 s]
[INFO] Apache Oozie MiniOozie ..... SUCCESS [ 9.647 s]
[INFO] Apache Oozie Distro ..... SUCCESS [ 27.966 s]
[INFO] Apache Oozie ZooKeeper Security Tests ..... SUCCESS [ 7.040 s]
[INFO] ------------------------------------------------------------------------
[INFO] BUILD SUCCESS
```

5，编译成功后在oozie-release-4.1.0/distro/target目录下，会生成如下的几个文件：

Java代码

```
drwxr-xr-x 2 root root 4096 5月 15 13:45 antrun
drwxr-xr-x 2 root root 4096 5月 15 13:45 archive-tmp
drwxr-xr-x 2 root root 4096 5月 15 13:45 maven-archiver
drwxr-xr-x 3 root root 4096 5月 15 13:46 oozie-4.1.0-distro
-rw-r--r-- 1 root root 201469924 5月 15 13:46 oozie-4.1.0-distro.tar.gz
-rw-r--r-- 1 root root 2875 5月 15 13:45 oozie-distro-4.1.0.jar
drwxr-xr-x 3 root root 4096 5月 15 13:45 tomcat
```

6，拷贝oozie-4.1.0-distro.tar.gz压缩包，至你需要安装的地方并解压，然后进入根目录下，

执行mkdir libext命令，创建libext目录

接着执行

cp ${HADOOP_HOME}/share/hadoop/*/*.jar libext/

cp ${HADOOP_HOME}/share/hadoop/*/lib/*.jar libext/

命令，将hadoop的相关的jar包拷贝至改目录

下载一个ext-2.2.zip包，也放入libext目录，由于oozie的js可能会依赖这个包，最新的版本应该不需要了，待验证？这个包，散仙在文末会上传到附件中，

7，删除libext下这几个包，因为会和hadoop的中的一些包冲突，造成类加载器无法识别重复的jsp，servlet或el解析器：

jasper-compiler-5.5.23.jar

jasper-runtime-5.5.23.jar

jsp-api-2.1.jar

8,修改conf/oozie-site.xml文件，更改以下几个地方：

Xml代码

```
<!-- 修改对应的hadoop的安装用户，散仙这里是search -->
<property>
<name>oozie.system.id</name>
<value>oozie-search</value>
<description>
The Oozie system ID.
</description>
</property>

<!-- 修改hadoop的conf的文件目录 -->
<property>
<name>oozie.service.HadoopAccessorService.hadoop.configurations</name>
<value>*=/home/search/hadoop/etc/hadoop</value>
<description>
Comma separated AUTHORITY=HADOOP_CONF_DIR, where AUTHORITY is the HOST:PORT of
the Hadoop service (JobTracker, HDFS). The wildcard '*' configuration is
used when there is no exact match for an authority. The HADOOP_CONF_DIR contains
the relevant Hadoop *-site.xml files. If the path is relative is looked within
the Oozie configuration directory; though the path can be absolute (i.e. to point
to Hadoop client conf/ directories in the local filesystem.
</description>
</property>

<!-- 修改oozie的share lib的HDFS目录 -->
<property>
<name>oozie.service.WorkflowAppService.system.libpath</name>
<value>/user/search/share/lib</value>
<description>
System library path to use for workflow applications.
This path is added to workflow application if their job properties sets
the property 'oozie.use.system.libpath' to true.
</description>
</property>
```

```
35.

36. <!-- 修改代理用户Hue需要用到，下面这两个配置，在Hadoop的core-site.xml中，同样需要添加，代理用户提交作业功能 -->
37. <property>
38. <name>oozie.service.ProxyUserService.proxyuser.search.hosts</name>
39. <value>*</value>
40. </property>
41.

42. <property>
43. <name>oozie.service.ProxyUserService.proxyuser.search.groups</name>
44. <value>*</value>
45. </property>
```

9，删除/home/search/oozie-4.1.0/conf/hadoop-conf下的core-site.xml文件，

将/home/search/hadoop/etc/hadoop/下的所有配置文件，拷贝到此处

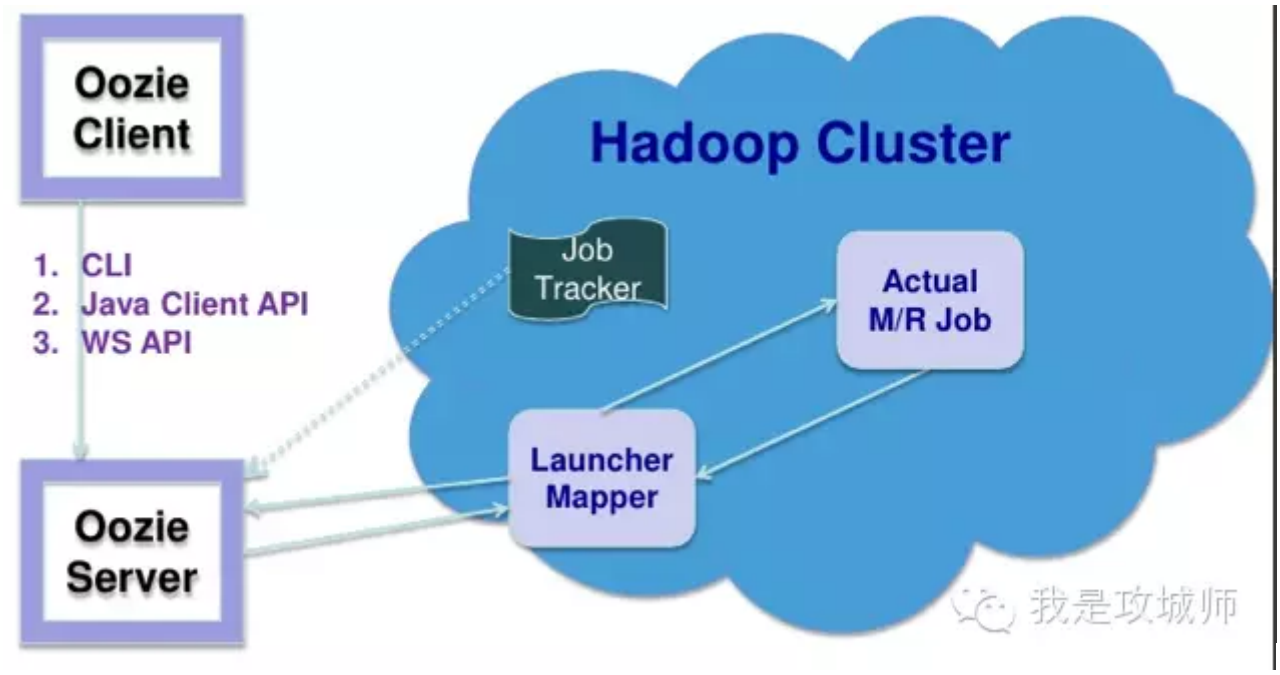
（6）执行bin/oozie-setup.sh prepare-war命令，重新生成war包

（7）执行bin/oozie-setup.sh sharelib create -fs hdfs://<namenode-hostname>:8020命令，将share下面的共享jar拷贝至HDFS中，

此处，也可以自己使用hadoop fs -copyFromLocal share/ /hdfs/xxx拷贝

（8）执行bin/oozie-setup.sh db create -run初始化oozie数据库

（9）执行bin/oozied.sh start启动oozie server

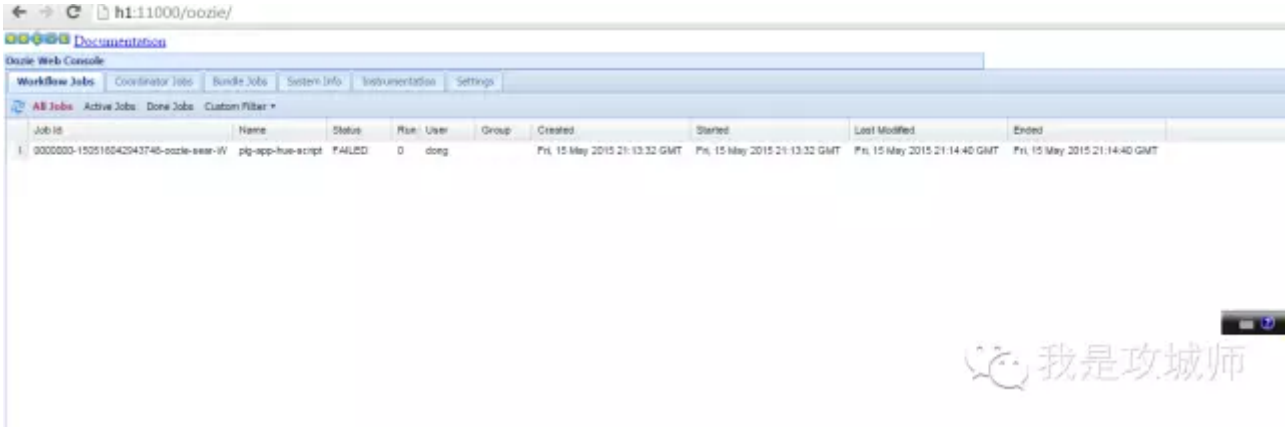


（10） 执行bin/oozie admin -oozie http://localhost:11000/oozie -status）返回Normal，即代表安装成功

Java代码

```
1. [search@h1 oozie-4.1.0]$ bin/oozie admin -oozie http://localhost:11000/oozie -status
2. System mode: NORMAL
3. [search@h1 oozie-4.1.0]$
```

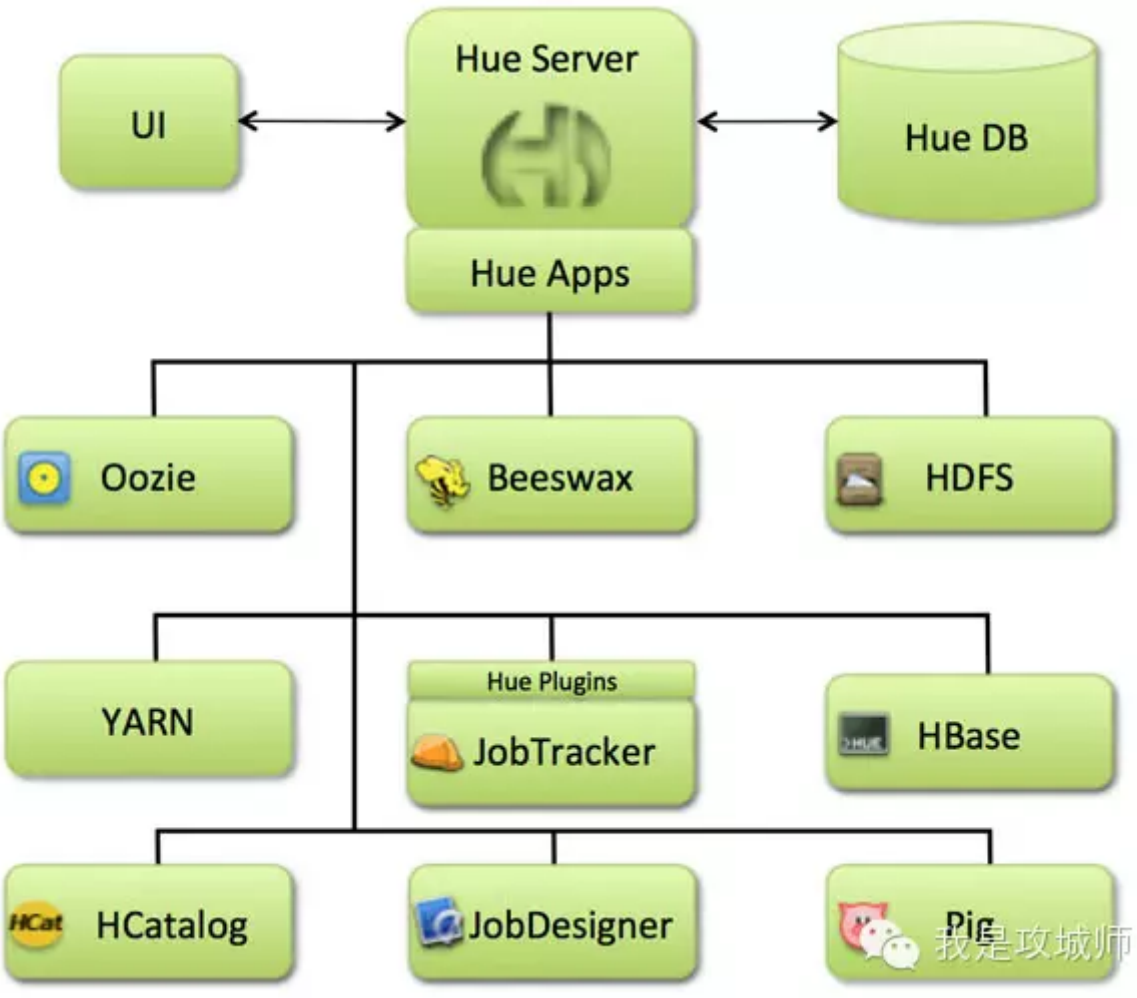
(11)在win上访问测试



(12)看到上图，说明你已经成功安装了，关系服务的命令

bin/oozied.sh stop,如果说不能停止，需要手动去删掉pid文件，然后在关闭。

oozie安装成功，很重要，因为Hue需要依赖它，做任务调度，下一篇文章，散仙就总结下hue安装笔记。


UI
Hue Server
Hue DB
Hue Apps
Oozie
Beeswax
HDFS
YARN
Hue Plugins
JobTracker
HBase
HCatalog
JobDesigner
Pig
我是攻城师

阅读原文