**Hadoop安装和使用**

**（安装hadoop和之前确保已安装java、ruby ）**

1：how to install hadoop

* (search all hadoop versions)

brew search hadoop

* (install hadoop version 1.2.1)

install homebrew/versions/hadoop121

* 进入(/usr/local/Cellar/hadoop121/1.2.1/libexec/conf/)

2： how to config hadoop

a.修改hadoop-env.sh(指定hadoop rvm 用于安装gem)

[[ -s "$HOME/.rvm/scripts/rvm" ]] && source "$HOME/.rvm/scripts/rvm"

rvm use [ruby-1.9.3-p551@default](mailto:ruby-1.9.3-p551@default)

b.core-site.xml配置HDFS master ip and port

<configuration>

<property>

<name>fs.default.name</name>

<value>hdfs://localhost:9000</value>

</property>

</configuration>

c．Mappred-site.xml配置HDFS MapReduce JobTracker ip and port

<configuration>

<property>

<name>mapred.job.tracker</name>

<value>localhost:9001</value>

</property>

</configuration>

3: **SSH无密码验证配置**

ssh-keygen -t rsa 会生成秘钥

cd .ssh/;ls

cat ~/.ssh/id\_rsa.pub >> ~/.ssh/authorized\_keys  查看秘钥拷贝

测试：ssh localhost

4:配置环境变量

加入：export HADOOP\_HOME=/usr/local/Cellar/hadoop121/1.2.1/libexec

使配置生效：source /etc/profile

如:wq 时报：E45：‘readonly’ option is set (add! To override)

可输入 :w !sudo tee % 进行保存

5:测试hadoop是否安装成功

执行jps 显示如下

96936 Jps

96124 JobTracker

94308 NameNode

95618 SecondaryNameNode

96777 TaskTracker

75987 RunJar

94962 DataNode

6：启动hadoop（/usr/local/Cellar/hadoop121/1.2.1/bin）

./start-all

7：

（注：如果执行出错 Unable to load realm info from SCDynamicStore

在hadoop-env.sh中加入

export HADOOP\_OPTS="-Djava.security.krb5.realm= -Djava.security.krb5.kdc="

在/usr/local/Cellar/hadoop121/1.2.1/libexec

export HADOOP\_HOME\_WARN\_SUPPRESS=1

如果输入： hadoop 执行出错 -bash: hadoop: command not found

在.bashrc中加入export PATH=$PATH:$HADOOP\_HOME/bin/ 并重启 source .bashrc）

**PIG安装和使用**

1: brew install pig

2：修改配置文件 /etc/profile

export HADOOP\_HOME=/usr/local/Cellar/hadoop121/1.2.1/libexec

export PIG\_CLASSPATH=$HADOOP\_HOME/conf

path=$JAVA\_HOME/bin:$HADOOP\_HOME/bin:$PIG\_CLASSPATH:$PATH

使配置生效：source /etc/profile

3：执行pig

出现grunt> 成功

4: 执行tracking-ingest 中得pig文件

**当运行pig脚本时，会将日志文件转化为Mapreduce之后hadoop会执行。**

**以app/click中得pig文件执行过程为例**：

1. 在执行该命令前请向hadoop中传入pig依赖的文件：

('/staging/tracking/lookups/GeoLiteCity.dat#GeoLiteCity.dat','/staging/tracking/lookups/superadmin.csv#superadmin.csv','/staging/tracking/lookups/location.csv#location.csv');

1. 将需要的文件放入hadoop中命令如下

hadoop fs –mkdir /staging/tracking/lookups

hadoop fs –put ~/Desktop/ GeoLiteCity.dat /staging/tracking/lookups

1. 将需要转化的日志文件存放在hadoop中命令如下

hadoop fs –mkdir /staging/tracking/incoming

hadoop fs –put

~/Desktop/ click.2015073006\_0.log /staging/tracking/incoming/click.2015073006\_0.log

1. 创建转化日志输出文件目录

hadoop fs –mkdir /staging/tracking/parsed

1. 进入pig文件的指定目录为pig文件传入2个参数，（需要执行的日志文件，输出文件）

命令如下

Pig -p INFILE=/staging/tracking/incoming/click.2015073006\_0.log -p OUTFILE=/staging/tracking/parsed/batch\_id=20150730060log click\_tracking.pig

（请注意输出日志文件夹名称规则batch\_id=infile的文件名去掉\_去掉. 这是因为之后会将转化后的文件存入hive中）

5：

（注：如果运行出错时报某个gem无法require 请确保hadoop-env中是否加入了

[[ -s "$HOME/.rvm/scripts/rvm" ]] && source "$HOME/.rvm/scripts/rvm"

rvm use [ruby-1.9.3-p551@default](mailto:ruby-1.9.3-p551@default)）

Hive的安装和使用

1：brew search hive(搜索出hive0.10.1)

2：brew install homebrew/versions/hive010

在安装过程中你会发现首先安装的是hadoop 2.X的版本不用理会，在下载完成之后在/use/local/Cell下删除hadoop 2.x的版本

3：配置环境

打开 /etc/procfile 加入以下命令

export HIVE\_HOME=/usr/local/Cellar/hive

export PATH=$PATH:$HIVE\_HOME/bin

export CLASSPATH=$CLASSPATH:$HIVE\_HOME/lib

4：执行hive

出现hive> 安装成功

如果执行hive命令如果出现以下错误

Unable to determine Hadoop version information.：'hadoop version' returned:

Using /Users/iclick/.rvm/gems/ruby-1.9.3-p551 Hadoop 1.2.1 Subversion https://svn.apache.org/repos/asf/hadoop/common/branches/branch-1.2 -r 1503152 Compiled by mattf on Mon Jul 22 15:23:09 PDT 2013 From source with checksum 6923c86528809c4e7e6f493b6b413a9a This command was run using /usr/local/Cellar/hadoop121/1.2.1/libexec/hadoop-core-1.2.1.jar

解决方法：在.profile中加入hadoop version：

export HADOOP\_VERSION=‘1.2.1'

然后source  .profile  在运行hive

如果已经添加了上面命令，仍然回报同样的错，重新执行即可

5：创建数据库和表

进入hive之后，按着tracking\_ingest/setup/creative\_hive\_objects.sql执行命令，在添加clicks\_parsed、adgroups\_parsed、imageviews\_parsed、eventdatas\_parsed、rtb\_datas\_parsed时，请确保存在hadoop 目录下存在sql文件中提到的目录

6：将fluent日志文件转化后的文件存储在hive中

执行tracking\_ingest/click/click\_hd\_to\_hive.sql,在执行

**alter table** clicks\_parsed **add** if **not exists** partition (batch\_id = **'${hiveconf:batch\_id}'**);

将**'${hiveconf:batch\_id}'换成output文件的名称（eg:** 20150730060log**）**

7:查看是否存入成功

进入hive中选中数据库，select \* from clicks\_parsed limit 1; 查看数据即可