第四十九章：Hive数据仓库工具

**一、Hive的概述；**

**二、Hive的体系架构；**

**三、探讨Hadoop全家（生态圈）；**

**四、Hive与传统数据库相比；**

**五、Hive的数据模型（自主学习）；**

**六、搭建Haoop 2.7.6结合hive2.3.3实现数据仓库管理；**

**一、Hive的概述；**

**背景：**Hive起源于Facebook（一个美国的社交服务网络）。Facebook有着大量的数据，而Hadoop是一个开源的MapReduce实现，可以轻松处理大量的数据。但是MapReduce程序对于Java程序员来说比较容易写，但是对于其他语言使用者来说不太方便。此时Facebook最早地开始研发Hive，它让对Hadoop使用SQL查询（实际上SQL后台转化为了MapReduce）成为可能，那些非Java程序员也可以更方便地使用。hive最早的目的也就是为了分析处理海量的日志；

**概述：**hive是基于Hadoop的一个数据仓库工具，可以将结构化的数据文件映射为一张数据库表，并提供简单的sql查询功能， Hadoop 通常都有较高的延迟并且在作业提交和调度的时候需要大量的开销。因此Hive并不能够在大规模数据集上实现低延迟快速的查询，例如，Hive 在几百MB 的数据集上执行查询一般有分钟级的时间延迟；

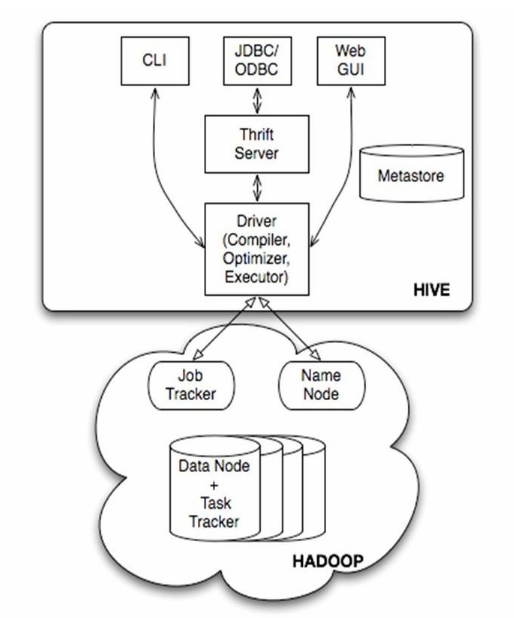


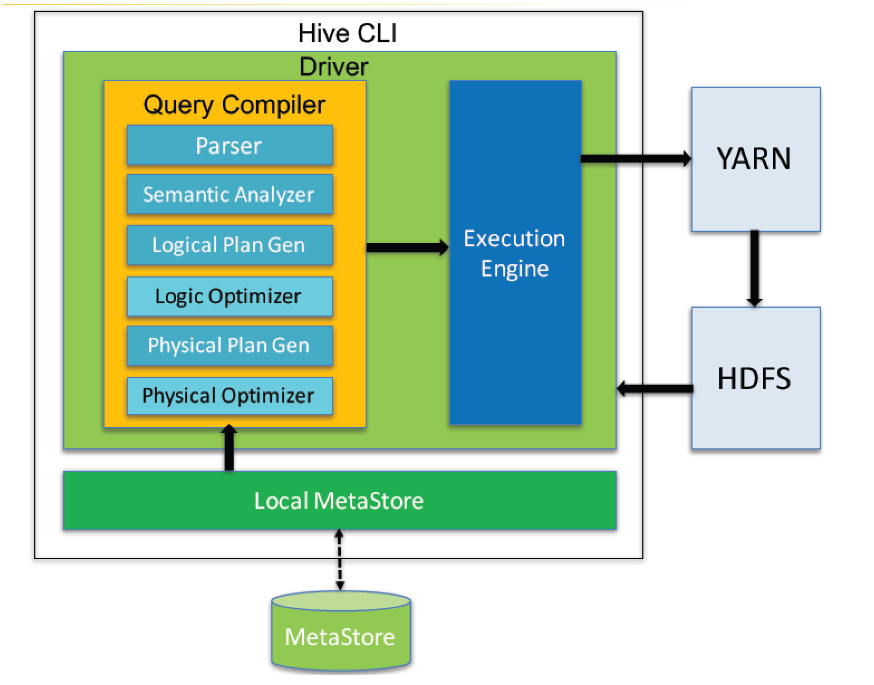
Hive可以将sql语句转换为MapReduce任务进行运行。其优点是学习成本低，可以通过类SQL语句快速实现简单的MapReduce统计，不必开发专门的MapReduce应用，十分适合数据仓库的统计分析；

**二、Hive的体系架构；**

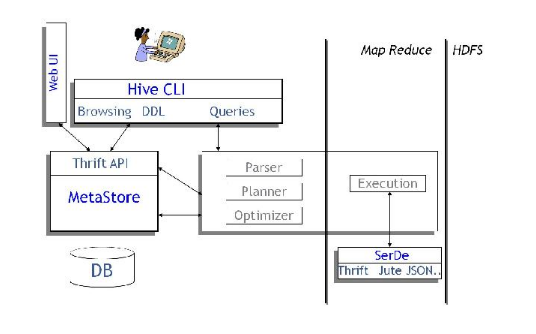
**体系架构：**

1.用户接口主要有三个：CLI，Client 和 WUI。其中最常用的是CLI，Cli启动的时候，会同时启动一个Hive副本。Client是Hive的客户端，用户连接至Hive Server。在启动 Client模式的时候，需要指出Hive Server所在节点，并且在该节点启动Hive Server。WUI是通过浏览器访问Hive；  
2.Hive将元数据存储在数据库中，如mysql、derby。Hive中的元数据包括表的名字，表的列和分区及其属性，表的属性（是否为外部表等），表的数据所在目录等；  
3.解释器、编译器、优化器完成HQL查询语句从词法分析、语法分析、编译、优化以及查询计划的生成。生成的查询计划存储在HDFS中，并在随后有MapReduce调用执行；  
4.Hive的数据存储在HDFS中，大部分的查询、计算由MapReduce完成；





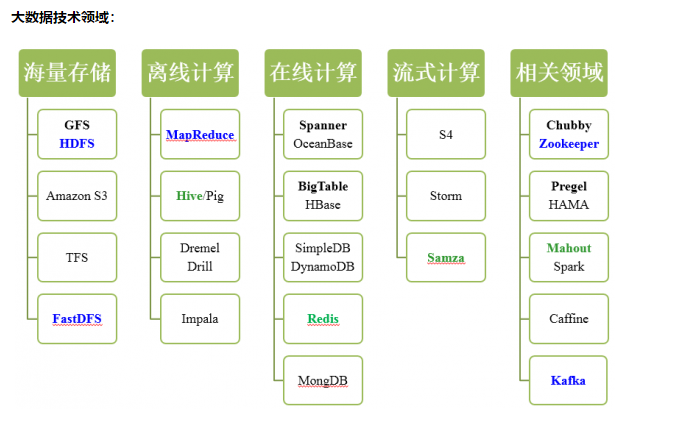
**Hive执行请求的流程：**

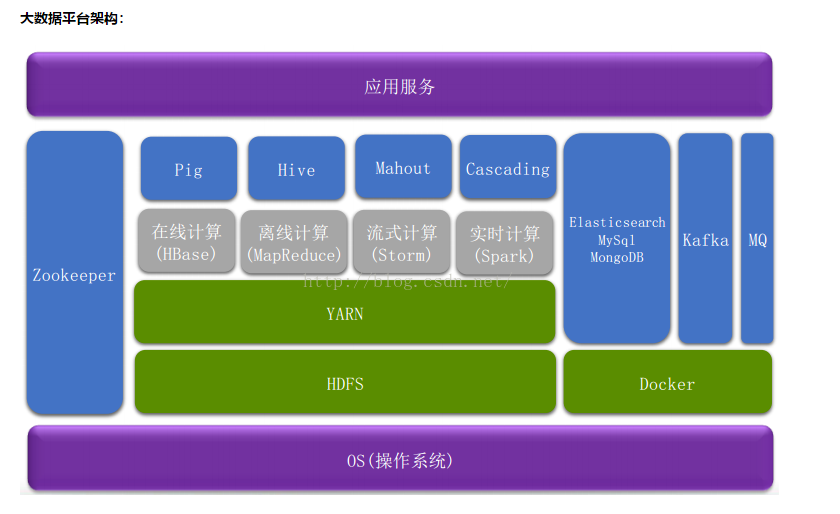
****

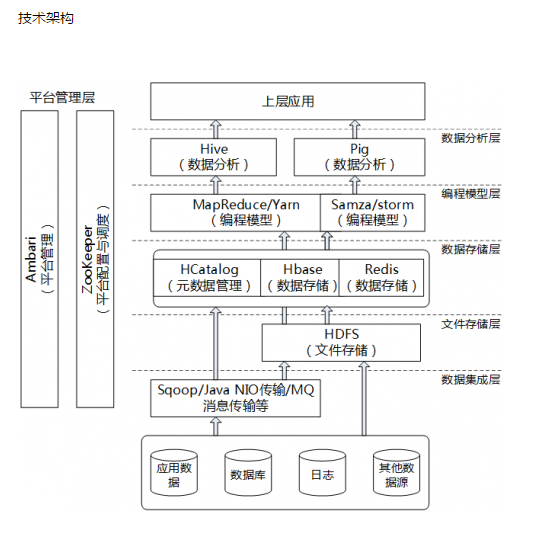
**三、探讨Hadoop全家（生态圈）；**

https://www.sohu.com/a/218420440\_505794 ##Hadoop全家

https://www.cnblogs.com/zhangwuji/p/7594725.html ##Hadoop发展史

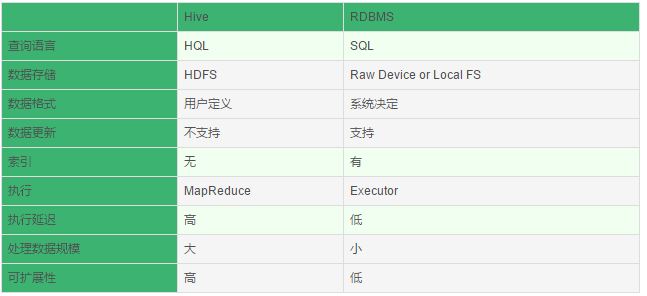






**四、Hive与传统数据库相比；**

 由于Hive采用了SQL的查询语言HQL，因此很容易将Hive理解为数据库。其实从结构上来看，Hive和数据库除了拥有类似的查询语言，再无类似之处。数据库可以用在Online的应用中，但是Hive是为数据仓库而设计的，清楚这一点，有助于从应用角度理解Hive的特性；



**五、Hive的数据模型（自主学习）；**

**概述：**对于数据存储，Hive没有专门的数据存储格式，也没有为数据建立索引，用户可以非常自由的组织Hive中的表，只需要在创建表的时候告诉Hive数据中的列分隔符和行分隔符，Hive就可以解析数据。Hive中所有的数据都存储在HDFS中，存储结构主要包括数据库、文件、表和视图。Hive中包含以下数据模型：Table内部表、External Table外部表、Partition分区、Bucket桶。自行百度Hive的数据模型；

**六、搭建Haoop 2.7.6结合hive2.3.3实现数据仓库管理；**

**案例环境：**

|  |  |  |  |
| --- | --- | --- | --- |
| 系统类型 | IP地址 | 主机名、角色 | 所需软件 |
| Centos 7.4 1708 64bit | 192.168.100.101 | master  hadoop：namenode  hive | hadoop-2.7.6.tar.gz  jdk-8u171-linux-x64.tar.gz  apache-hive-2.3.3-bin.tar.gz |
| Centos 7.4 1708 64bit | 192.168.100.102 | slave1  hadoop：datanode | hadoop-2.7.6.tar.gz  jdk-8u171-linux-x64.tar.gz |
| Centos 7.4 1708 64bit | 192.168.100.103 | slave2  hadoop：datanode | hadoop-2.7.6.tar.gz  jdk-8u171-linux-x64.tar.gz |

**案例步骤：**

* 安装hive软件程序；
* 准备hive的元数据存储位置Mysql数据库；
* 准备hive服务需要存储文件的目录；
* 配置hive的配置文件；
* 配置hive服务连接mysql数据库的驱动；
* 初始化hive连接mysql数据库；
* 启动并连接hive服务；
* 学习hive中的管理操作；
* 将文件内容导入到hive的表中；
* 通过web页面查看HDFS文件系统中的数据；
* 通过mysql管理控制台验证mysql中存储的hive的元数据；
* **安装hive软件程序；**

[root@master ~]# ls /root/apache-hive-2.3.3-bin.tar.gz

/root/apache-hive-2.3.3-bin.tar.gz

[root@master ~]# tar zxvf /root/apache-hive-2.3.3-bin.tar.gz

[root@master ~]# mv apache-hive-2.3.3-bin/ /usr/local/hive

[root@master ~]# chown hadoop:hadoop /usr/local/hive/ -R

[root@master ~]# vi /etc/profile

export HIVE\_HOME=/usr/local/hive

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

export CLASSPATH=$CLASSPATH:/usr/local/hive/lib/\*:.

[root@master ~]# source /etc/profile

* **准备hive的元数据存储位置Mysql数据库；**

[root@master ~]# yum -y install mariadb-server mariadb

[root@master ~]# sed -i 's/\[mysqld\]/\[mysqld\]\nskip\_name\_resolve/g' /etc/my.cnf

##注：需要关闭mysql服务器的反向解析，不然会在初始化myssql数据库时出现反向解析情况，无法初始化

[root@master ~]# systemctl start mariadb

[root@master ~]# systemctl enable mariadb

Created symlink from /etc/systemd/system/multi-user.target.wants/mariadb.service to /usr/lib/systemd/system/mariadb.service.

[root@master ~]# mysqladmin -uroot password

[root@master ~]# mysql -uroot -p123123

MariaDB [(none)]> grant all on \*.\* to root@'192.168.100.101' identified by '123123';

MariaDB [(none)]> flush privileges;

MariaDB [(none)]> exit

* **准备hive服务需要存储文件的目录；**

[root@master ~]# su - hadoop

[hadoop@master ~]$ hadoop fs -mkdir -p /user/hive/warehouse

[hadoop@master ~]$ hadoop fs -chmod 777 /user/hive/warehouse

[hadoop@master ~]$ hadoop fs -mkdir -p /tmp/hive/

[hadoop@master ~]$ hadoop fs -chmod 777 /tmp/hive

[hadoop@master ~]$ hadoop fs -ls /

Found 2 items

drwxr-xr-x - hadoop supergroup 0 2018-08-27 20:51 /tmp

drwxr-xr-x - hadoop supergroup 0 2018-08-27 20:51 /user

[hadoop@master ~]$ mkdir /usr/local/hive/tmp

[hadoop@master ~]$ mkdir /usr/local/hive/tmp/root

* **配置hive的配置文件；**

[hadoop@master ~]$ cp /usr/local/hive/conf/hive-default.xml.template /usr/local/hive/conf/hive-site.xml

[hadoop@master ~]$ mv /usr/local/hive/conf/hive-env.sh.template /usr/local/hive/conf/hive-env.sh

[hadoop@master ~]$ vi /usr/local/hive/conf/hive-site.xml

364 <property>

365 <name>hive.metastore.warehouse.dir</name>

366 <value>/user/hive/warehouse</value>

367 <description>指定hdfs上存放hive的实体数据位置</description>

368 </property>

37 <property>

38 <name>hive.exec.scratchdir</name>

39 <value>/tmp/hive</value>

40 <description>指定hdfs上存储hive的临时文件的位置</description>

41 </property>

73 <property>

74 <name>hive.exec.local.scratchdir</name>

75 <value>/usr/local/hive/tmp/root/</value>

76 <description>指定hive本地服务器的本地目录</description>

77 </property>

78 <property>

79 <name>hive.downloaded.resources.dir</name>

80 <value>/usr/local/hive/tmp/${hive.session.id}\_resources</value>

81 <description>指定hive本地服务器的session会话文件的存储位置</description>

82 </property>

1683 <property>

1684 <name>hive.querylog.location</name>

1685 <value>/usr/local/hive/tmp/root/</value>

1686 <description>指定hive本地服务器的日志文件存放位置</description>

1687 </property>

3974 <property>

3975 <name>hive.server2.logging.operation.log.location</name>

3976 <value>/usr/local/hive/tmp/root/operation\_logs</value>

3977 <description>指定hive本地服务器的日志文件存放位置</description>

3978 </property>

543 <property>

544 <name>javax.jdo.option.ConnectionURL</name>

545 <value>jdbc:mysql://192.168.100.101:3306/hive?createDatabaseIfNotExist=true</value>

546 <description>指定hive的元数据所存储的mysql的ip、端口、数据库名

550 </description>

551 </property>

1018 <property>

1019 <name>javax.jdo.option.ConnectionDriverName</name>

1020 <value>com.mysql.jdbc.Driver</value>

1021 <description>指定连接mysql数据库的驱动类型</description>

1022 </property>

1043 <property>

1044 <name>javax.jdo.option.ConnectionUserName</name>

1045 <value>root</value>

1046 <description>指定连接mysql数据库的用户名</description>

1047 </property>

528 <property>

529 <name>javax.jdo.option.ConnectionPassword</name>

530 <value>123123</value>

531 <description>指定连接mysql数据库的密码</description>

532 </property>

744 <property>

745 <name>hive.metastore.schema.verification</name>

746 <value>false</value>

747 <description>关闭元数据的文件检查

753 </description>

754 </property>

:wq

[hadoop@master ~]$ vi /usr/local/hive/conf/hive-env.sh ##设置hive的环境变量

export HADOOP\_HOME=/usr/local/hadoop

export HIVE\_CONF\_DIR=/usr/local/hive/conf

export HIVE\_AUX\_JARS\_PATH=/usr/local/hive/lib

:wq

* **配置hive服务连接mysql数据库的驱动；**

[hadoop@master ~]$ wget https://cdn.mysql.com//Downloads/Connector-J/mysql-connector-java-5.1.46.tar.gz

[hadoop@master ~]$ ls mysql-connector-java-5.1.46.tar.gz

mysql-connector-java-5.1.46.tar.gz

[hadoop@master ~]$ tar zxvf mysql-connector-java-5.1.46.tar.gz

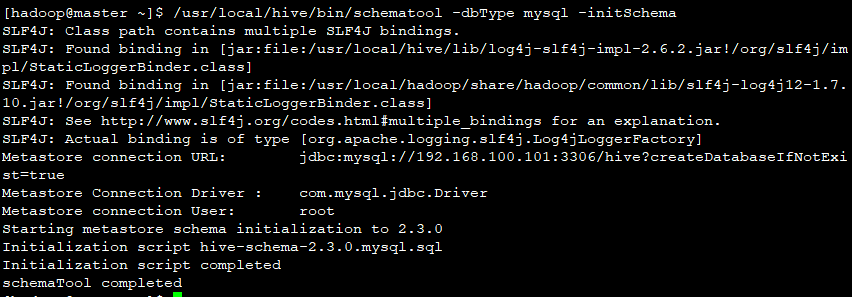
[hadoop@master ~]$ cp /home/hadoop/mysql-connector-java-5.1.46/mysql-connector-java-5.1.46.jar /usr/local/hive/lib/

[hadoop@master ~]$ ls /usr/local/hive/lib/mysql-connector-java-5.1.46.jar

/usr/local/hive/lib/mysql-connector-java-5.1.46.jar

* **初始化hive连接mysql数据库；**

[hadoop@master ~]$ /usr/local/hive/bin/schematool -dbType mysql -initSchema



* **启动并连接hive服务；**

[hadoop@master ~]$ hive

hive> show databases;

OK

default

Time taken: 4.695 seconds, Fetched: 1 row(s)

* **学习hive中的管理操作；**

hive> show functions; ##查看可用函数

OK

!

!=

$sum0

%

&

\*

+

-

/

<

<=

...

hive> desc function sum; ##查看函数详细信息

OK

sum(x) - Returns the sum of a set of numbers

Time taken: 0.006 seconds, Fetched: 1 row(s)

hive> create database linuxfan; ##创建数据库linuxfan

OK

Time taken: 0.282 seconds

hive> show databases;

OK

default

linuxfan

Time taken: 0.023 seconds, Fetched: 2 row(s)

hive> use linuxfan;

OK

Time taken: 0.025 seconds

hive> create table t1(id int,name string) row format delimited fields terminated by '\t';

##创建表并且指定字段id和字段name之间的分隔使用\t制表符；

OK

Time taken: 0.581 seconds

hive> show tables;

OK

t1

Time taken: 0.042 seconds, Fetched: 1 row(s)

hive> exit;

* **将文件内容导入到hive的表中；**

[hadoop@master ~]$vi ./1.txt ##001和zs之间必须手动敲tab键，不得复制粘贴

001 zs

002 ls

003 ww

004 xm

[hadoop@master ~]$ hive

hive> load data local inpath '/home/hadoop/1.txt' into table linuxfan.t1;

Loading data to table linuxfan.t1;

OK

Time taken: 6.277 seconds

hive> select \* from linuxfan.t1;

OK

1 zs

2 ls

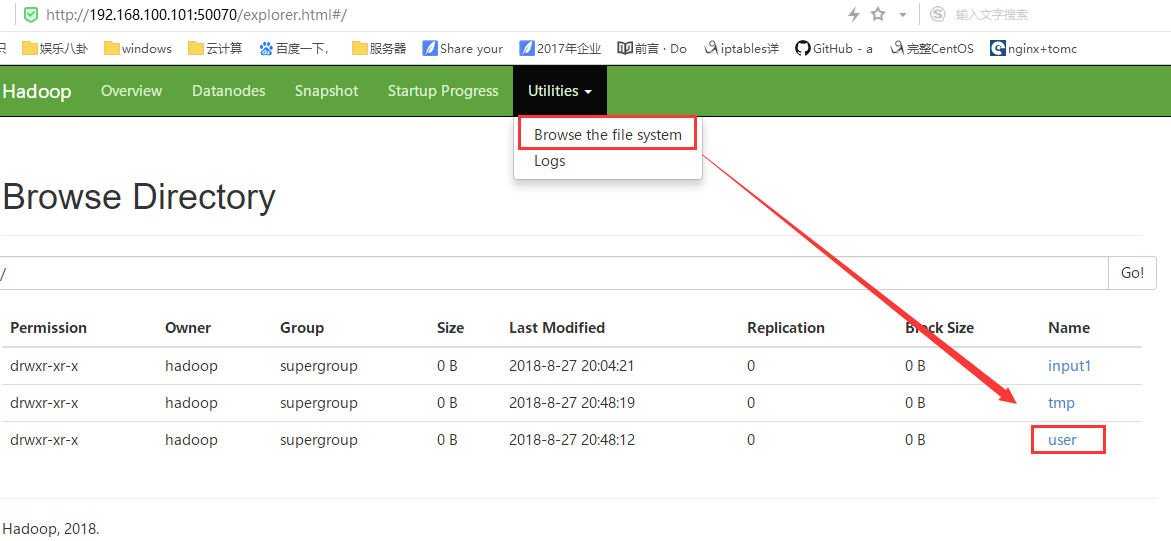
3 ww

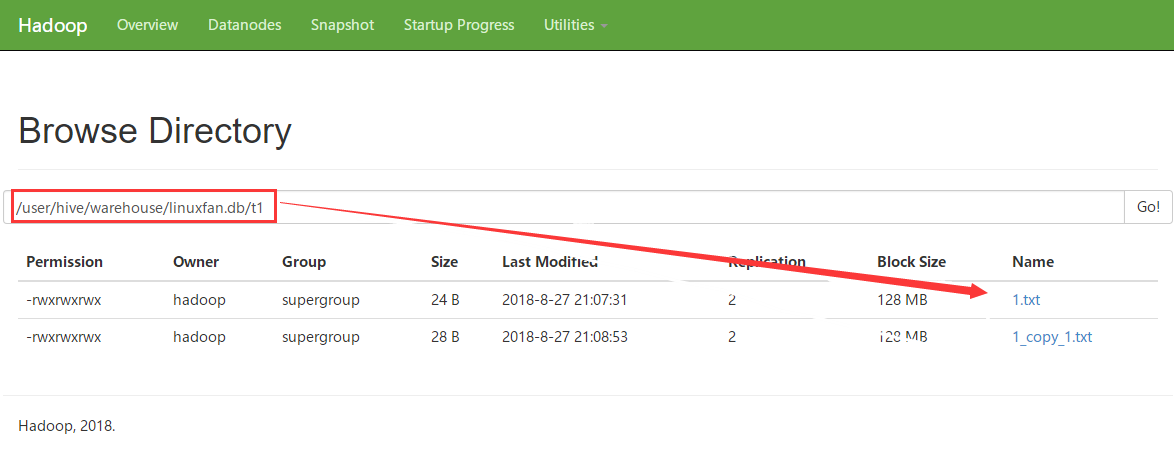
4 xm

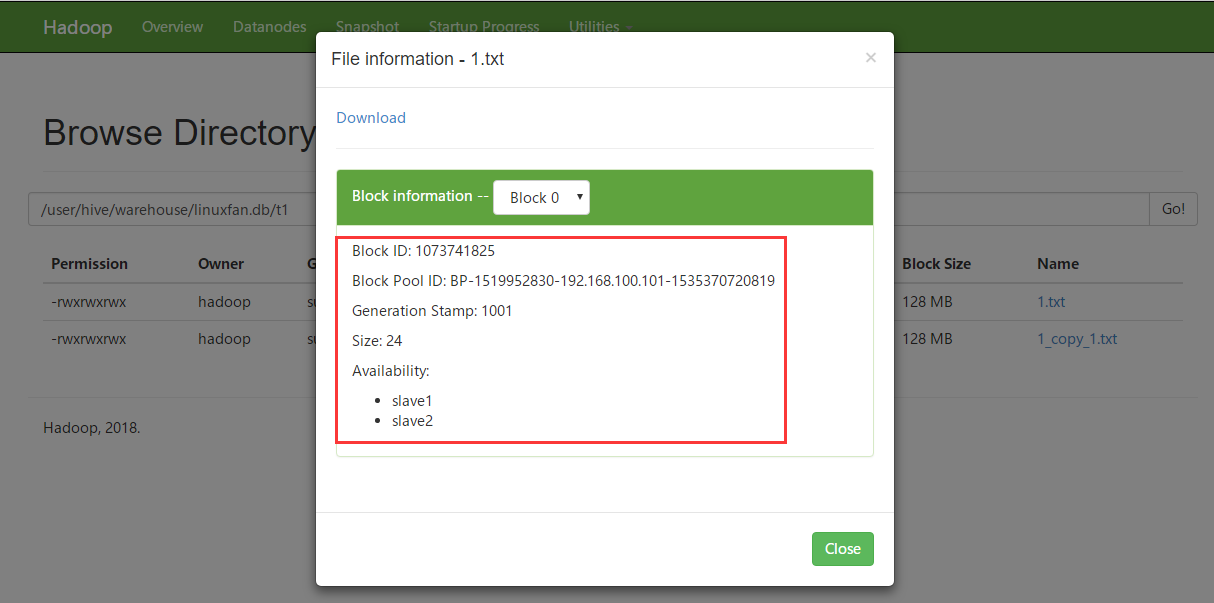
Time taken: 1.861 seconds, Fetched: 4 row(s)

hive> exit;

* **通过web页面查看HDFS文件系统中的数据；**







* **通过mysql管理控制台验证mysql中存储的hive的元数据；**

