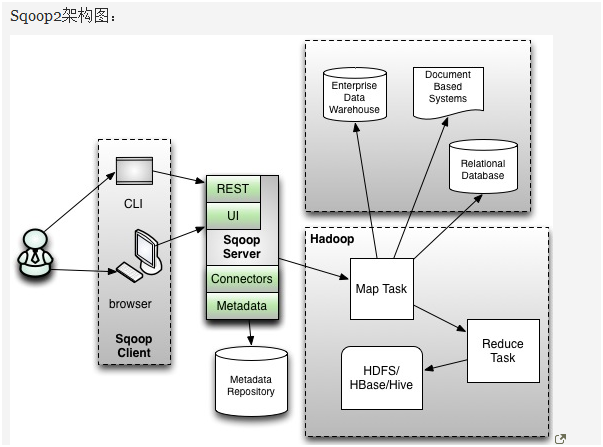
# Hdoop-Sqoop搭建

[mysql-connector-java.jar](/download/attachments/5932651/mysql-connector-java.jar?version=1&modificationDate=1503991075000&api=v2)**（sqoop连接mysql的jar包）**

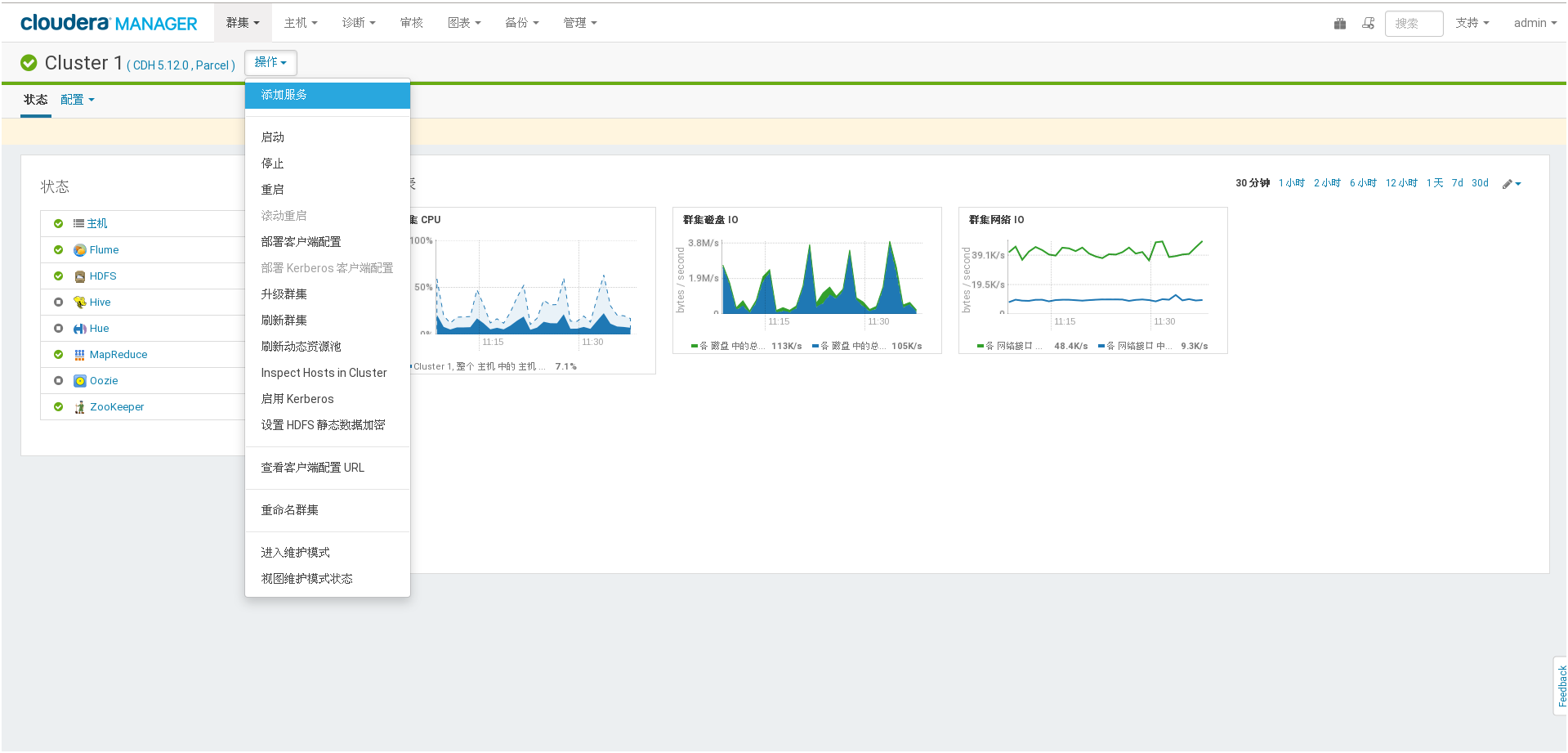
**一.**Sqoop简介

Sqoop是一个用来将Hadoop和关系型数据库中的数据相互转移的工具，可以将一个关系型数据库（例如 ： MySQL ,Oracle ,Postgres等）中的数据导进到Hadoop的HDFS中，也可以将HDFS的数据导进到关系型数据库中。

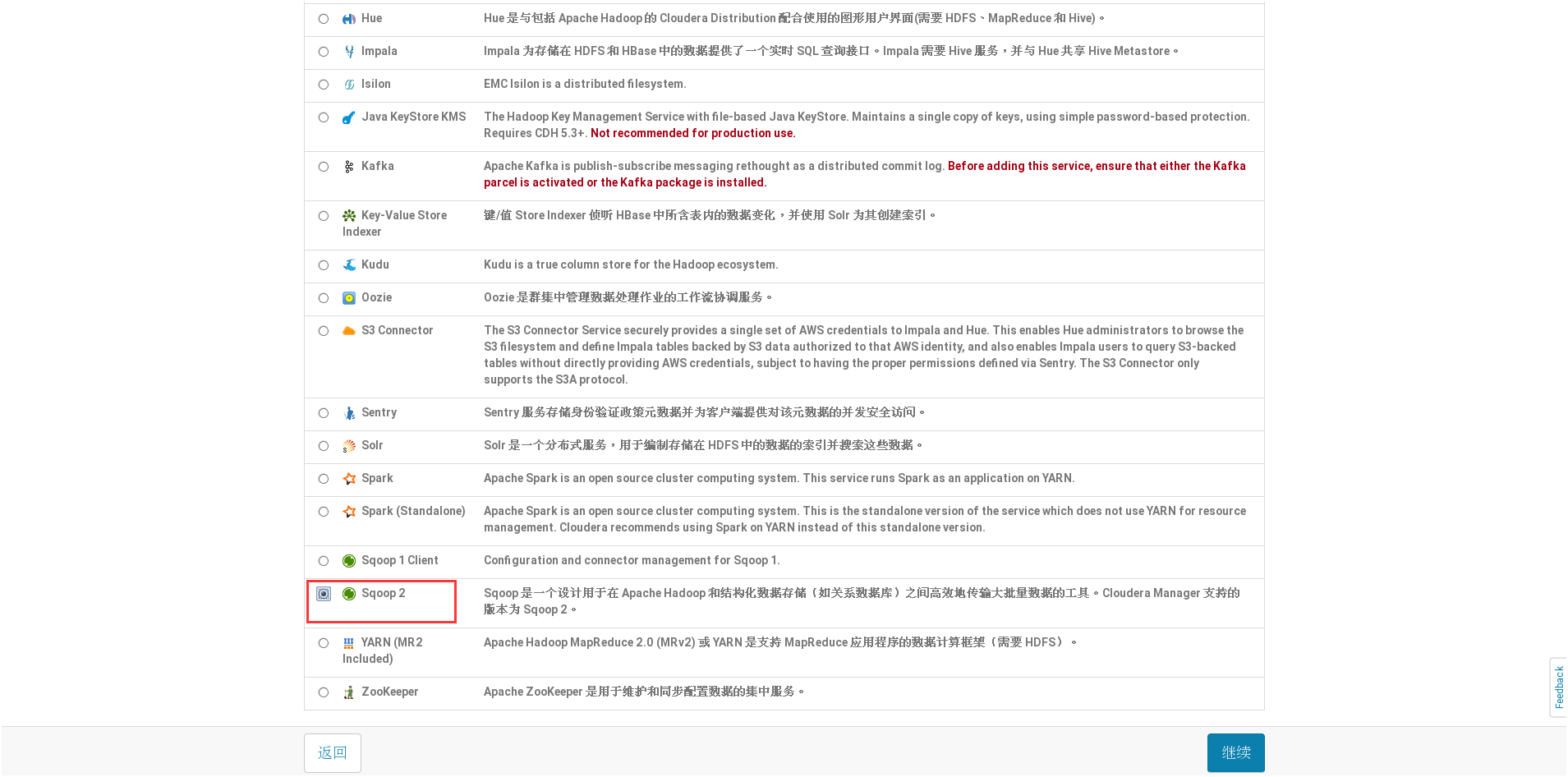


二.Sqoop安装（这里使用web界面安装）

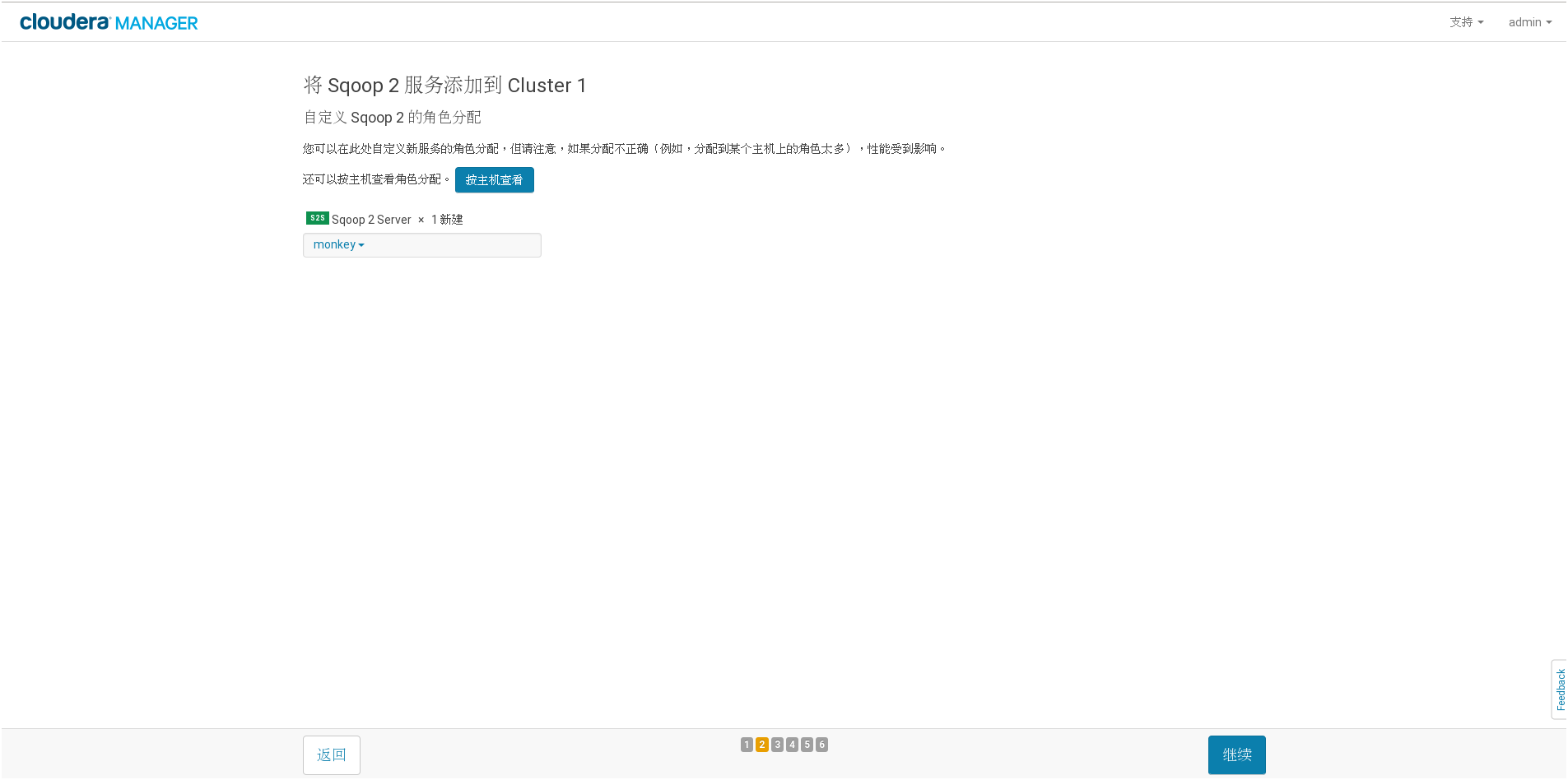
1.点击操作--添加服务



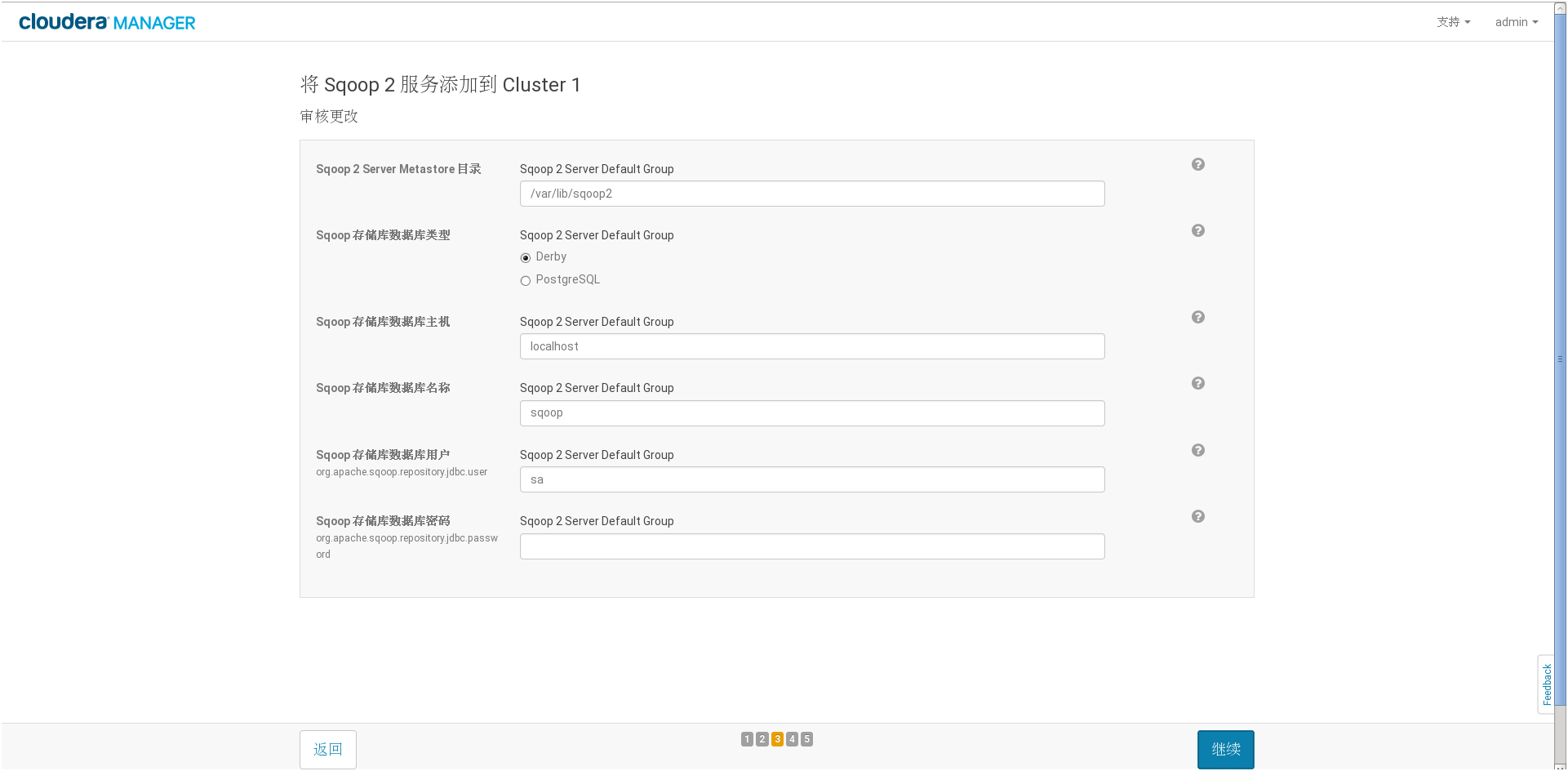
2.选择sqoop2，点击继续



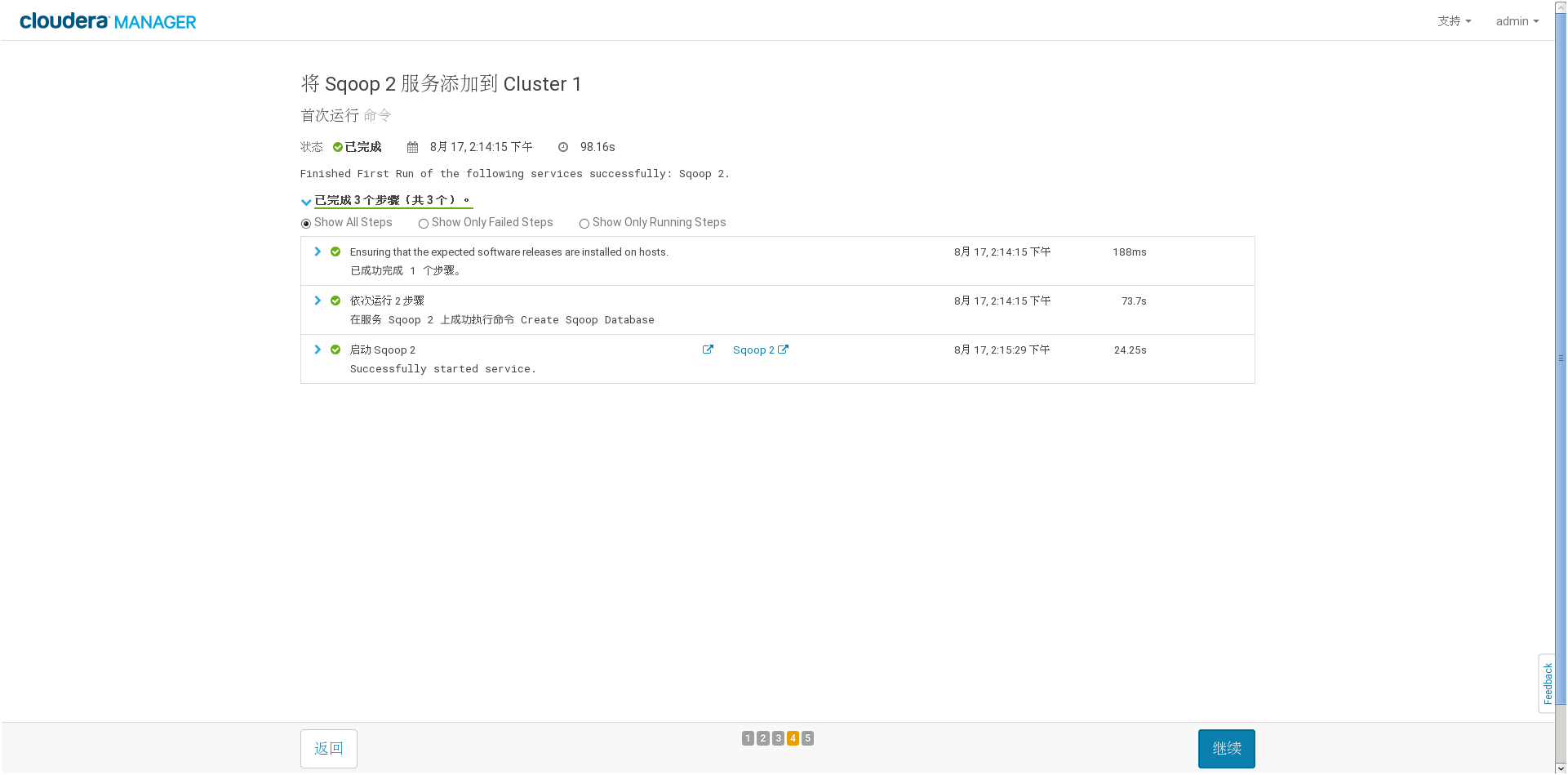
3.点击要选择安装选择Sqoop的节点，点击继续（不建议安装在CM管理节点上）



4.sqoop选项，默认点击继续



5.等待CM后台安装完Sqoop后，点击继续



6.Sqoop服务添加完，点击完成

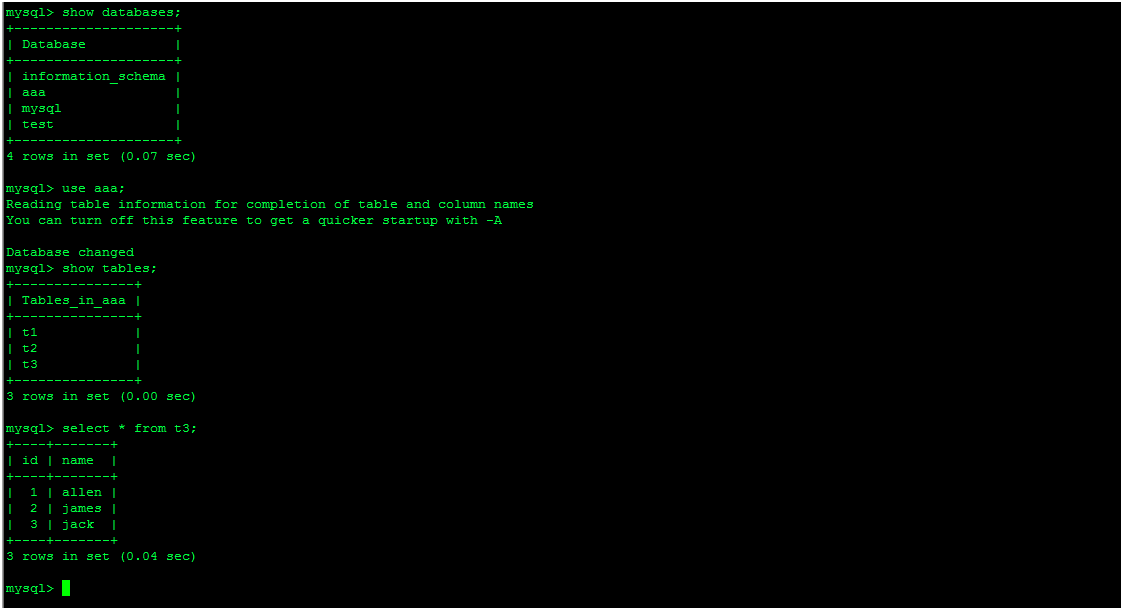


7.返回CM管理界面查看Sqoop服务是否启动



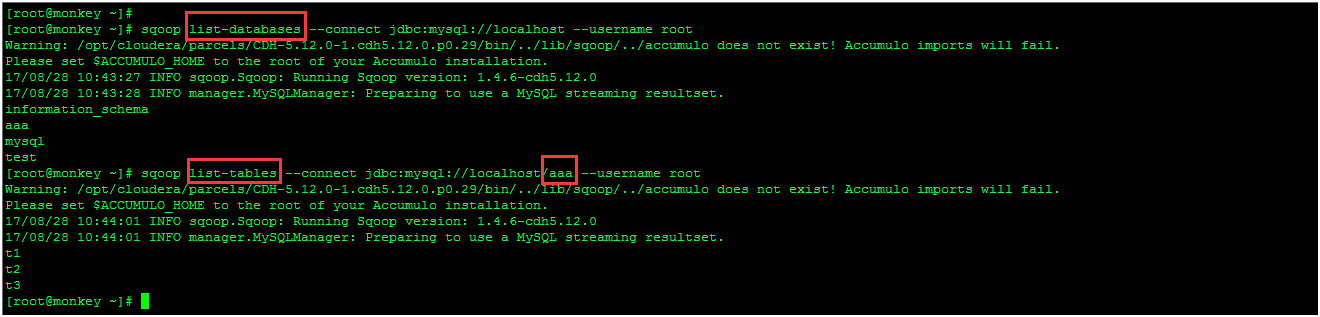
三.用Sqoop导入数据到HDFS

1.进入mysql，查看数据库表信息



2. 创建数据库链接：ln -s /usr/share/java/mysql-connector-java.jar /usr/lib/sqoop/lib/（注mysql-connector-java.jar此步骤较为关键如果没配置好无法连接进数据库,jar包可到hadoop官网下载）

3. 用sqoop连接数据库和表



4.用Sqoop导入数据到HDFS

[root@monkey ~]# sudo -u hdfs sqoop import --connect jdbc:mysql://localhost/aaa --table t3 --fields-terminated-by '\t' --username root

Warning: /opt/cloudera/parcels/CDH-5.12.0-1.cdh5.12.0.p0.29/bin/../lib/sqoop/../accumulo does not exist! Accumulo imports will fail.

Please set $ACCUMULO\_HOME to the root of your Accumulo installation.（此警告可忽略，或者注释掉bin/configure-sqoop文件中的以下部分

--------------------------------------------------

## Moved to be a runtime check in sqoop.

#if [ ! -d "${ $ACCUMULO\_HOME }" ];then

#  echo "Warning: $ $ACCUMULO\_HOME  does not exist! Accumulo imports will fail."

#  echo 'Please set $ $ACCUMULO\_HOME  to the root of your Accumulo installation.'#fi）

 ------------------------------------------------------

17/08/22 17:19:13 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.12.0

17/08/22 17:19:13 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.

17/08/22 17:19:13 INFO tool.CodeGenTool: Beginning code generation

17/08/22 17:19:13 INFO manager.SqlManager: Executing SQL statement: SELECT t.\* FROM `t3` AS t LIMIT 1

17/08/22 17:19:13 INFO manager.SqlManager: Executing SQL statement: SELECT t.\* FROM `t3` AS t LIMIT 1

17/08/22 17:19:13 INFO orm.CompilationManager: HADOOP\_MAPRED\_HOME is /opt/cloudera/parcels/CDH/lib/hadoop-0.20-mapreduce

注: /tmp/sqoop-hdfs/compile/562904808de81563ccd5c52404e233bc/t3.java使用或覆盖了已过时的API。

注: 有关详细信息, 请使用 -Xlint:deprecation 重新编译。

17/08/22 17:19:15 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-hdfs/compile/562904808de81563ccd5c52404e233bc/t3.jar

17/08/22 17:19:15 WARN manager.MySQLManager: It looks like you are importing from mysql.

17/08/22 17:19:15 WARN manager.MySQLManager: This transfer can be faster! Use the --direct

17/08/22 17:19:15 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.

17/08/22 17:19:15 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)

17/08/22 17:19:15 INFO mapreduce.ImportJobBase: Beginning import of t3

17/08/22 17:19:20 INFO db.DBInputFormat: Using read commited transaction isolation

17/08/22 17:19:20 INFO db.DataDrivenDBInputFormat: BoundingValsQuery: SELECT MIN(`id`), MAX(`id`) FROM `t3`

17/08/22 17:19:20 INFO db.IntegerSplitter: Split size: 0; Num splits: 4 from: 1 to: 3

17/08/22 17:19:20 INFO mapred.JobClient: Running job: job\_201708172318\_0004

17/08/22 17:19:21 INFO mapred.JobClient:  map 0% reduce 0%

17/08/22 17:19:53 INFO mapred.JobClient:  map 33% reduce 0%

17/08/22 17:20:22 INFO mapred.JobClient: Task Id : attempt\_201708172318\_0004\_m\_000000\_0, Status : FAILED---（信息显示mapreduce状态为failed 刚开始还以为我的mapreduce任务跑失败了，在观察了整个mapreduce任务后map还是跑完100%，看了下面的java报错

Communications link failure通信链路故障，怀疑各节点间网络连接导致或者建立的mysql的链接失效？？？，小伙伴们可以一起讨论分析mysql-connector-java.jar这个jar包代码）

java.lang.RuntimeException: java.lang.RuntimeException: com.mysql.jdbc.exceptions.jdbc4.CommunicationsException: Communications link failure

The last packet sent successfully to the server was 0 milliseconds ago. The driver has not received any packets from the server.

         at org.apache.sqoop.mapreduce.db.DBInputFormat.setDbConf(DBInputFormat.java:170)

         at org.apache.sqoop.mapreduce.db.DBInputFormat.setConf(DBInputFormat.java:161)

         at org.apache.hadoop.util.ReflectionUtils.setConf(ReflectionUtils.java:73)

         at org.apache.hadoop.util.ReflectionUtils.newInstance(ReflectionUtils.java:133)

         at org.apache.hadoop.mapred.MapTask.runNewMapper(MapTask.java:635)

         at org.apache.hadoop.mapred.MapTask.run(MapTask.java:330)

         at org.apache.hadoop.mapred.Child$4.run(Child.java:268)

         at java.security.AccessController.doPrivileged(Native Method)

         at javax.security.auth.Subject.doAs(Subject.java:415)

         at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1917)

         at org.apache.hadoop.mapred.Child.main(Chi

17/08/22 17:20:28 INFO mapred.JobClient:  map 67% reduce 0%

17/08/22 17:20:35 INFO mapred.JobClient: Task Id : attempt\_201708172318\_0004\_m\_000001\_0, Status : FAILED

java.lang.RuntimeException: java.lang.RuntimeException: com.mysql.jdbc.exceptions.jdbc4.CommunicationsException: Communications link failure

The last packet sent successfully to the server was 0 milliseconds ago. The driver has not received any packets from the server.

         at org.apache.sqoop.mapreduce.db.DBInputFormat.setDbConf(DBInputFormat.java:170)

         at org.apache.sqoop.mapreduce.db.DBInputFormat.setConf(DBInputFormat.java:161)

         at org.apache.hadoop.util.ReflectionUtils.setConf(ReflectionUtils.java:73)

         at org.apache.hadoop.util.ReflectionUtils.newInstance(ReflectionUtils.java:133)

         at org.apache.hadoop.mapred.MapTask.runNewMapper(MapTask.java:635)

         at org.apache.hadoop.mapred.MapTask.run(MapTask.java:330)

         at org.apache.hadoop.mapred.Child$4.run(Child.java:268)

         at java.security.AccessController.doPrivileged(Native Method)

         at javax.security.auth.Subject.doAs(Subject.java:415)

         at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1917)

         at org.apache.hadoop.mapred.Child.main(Chi

17/08/22 17:20:42 INFO mapred.JobClient:  map 100% reduce 0%----（在最后map跑完100%后，没有出现状态失败的提示？？？小伙伴们可以一起讨论分析下）

17/08/22 17:20:44 INFO mapred.JobClient: Job complete: job\_201708172318\_0004

17/08/22 17:20:44 INFO mapred.JobClient: Counters: 23

17/08/22 17:20:44 INFO mapred.JobClient:   File System Counters

17/08/22 17:20:44 INFO mapred.JobClient:     FILE: Number of bytes read=0

17/08/22 17:20:44 INFO mapred.JobClient:     FILE: Number of bytes written=838200

17/08/22 17:20:44 INFO mapred.JobClient:     FILE: Number of read operations=0

17/08/22 17:20:44 INFO mapred.JobClient:     FILE: Number of large read operations=0

17/08/22 17:20:44 INFO mapred.JobClient:     FILE: Number of write operations=0

17/08/22 17:20:44 INFO mapred.JobClient:     HDFS: Number of bytes read=295

17/08/22 17:20:44 INFO mapred.JobClient:     HDFS: Number of bytes written=23

17/08/22 17:20:44 INFO mapred.JobClient:     HDFS: Number of read operations=3

17/08/22 17:20:44 INFO mapred.JobClient:     HDFS: Number of large read operations=0

17/08/22 17:20:44 INFO mapred.JobClient:     HDFS: Number of write operations=3

17/08/22 17:20:44 INFO mapred.JobClient:   Job Counters

17/08/22 17:20:44 INFO mapred.JobClient:     Launched map tasks=5

17/08/22 17:20:44 INFO mapred.JobClient:     Total time spent by all maps in occupied slots (ms)=121897

17/08/22 17:20:44 INFO mapred.JobClient:     Total time spent by all reduces in occupied slots (ms)=0

17/08/22 17:20:44 INFO mapred.JobClient:     Total time spent by all maps waiting after reserving slots (ms)=0

17/08/22 17:20:44 INFO mapred.JobClient:     Total time spent by all reduces waiting after reserving slots (ms)=0

17/08/22 17:20:44 INFO mapred.JobClient:   Map-Reduce Framework

17/08/22 17:20:44 INFO mapred.JobClient:     Map input records=3

17/08/22 17:20:44 INFO mapred.JobClient:     Map output records=3

17/08/22 17:20:44 INFO mapred.JobClient:     Input split bytes=295

17/08/22 17:20:44 INFO mapred.JobClient:     Spilled Records=0

17/08/22 17:20:44 INFO mapred.JobClient:     CPU time spent (ms)=6720

17/08/22 17:20:44 INFO mapred.JobClient:     Physical memory (bytes) snapshot=547364864

17/08/22 17:20:44 INFO mapred.JobClient:     Virtual memory (bytes) snapshot=5017960448

17/08/22 17:20:44 INFO mapred.JobClient:     Total committed heap usage (bytes)=481296384

17/08/22 17:20:44 INFO mapreduce.ImportJobBase: Transferred 23 bytes in 87.9648 seconds (0.2615 bytes/sec)

17/08/22 17:20:44 INFO mapreduce.ImportJobBase: Retrieved 3 records.

5.查看导入到HDFS上的mysql数据信息（对比HDFS上的数据和mysql上aaa库上t3表信息）

