## Impala与 Hbase整合使用说明

目录

[Impala与 Hbase整合使用说明 1](#_Toc420055145)

[一、UDH环境集成配置 1](#_Toc420055146)

[二、Hive 映射已存在的Hbase表 2](#_Toc420055147)

[三、使用impala验证映射的Hbase表 2](#_Toc420055148)

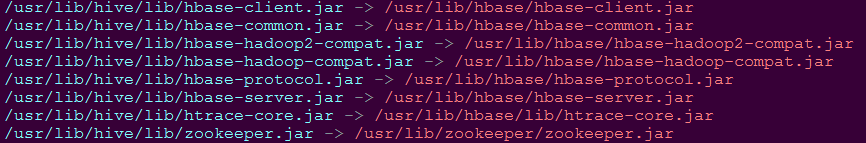
[四、常见问题汇总 3](#_Toc420055149)

HBase是一个基于列的NoSQL数据库，它可以实现的数据的灵活存储。它本身是一个大表，在一些应用中，通过设计RowKey，可以实现对海量数据的快速存储和访问。但是，对于复杂的查询统计类需求，如果直接基于HBase API来实现，性能非常差，或者，可以通过实现MapReduce程序来进行查询分析，这也继承了MapReduce所具备的延迟性。

实现Hive与HBase集成，我们知道，impala与hive共享相同的metstore数据库，Hive表的列的映射信息也适用impala，所以二者集成，可以使用我们熟悉的SQL,像操作传统关系型数据库一样，很容易给出复杂查询、统计分析的SQL设计，并通过Impala查询统计分析，比原生的MapReduce以及Hive的执行速度快很多。

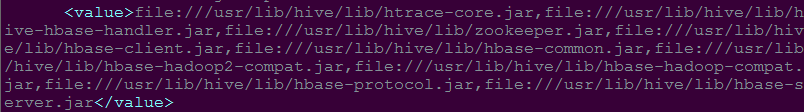
### 一、**UDH**环境**集成**配置

1.首先要确保/usr/lib/hive/lib 下HBase的jar包的版本要和实际环境中HBase的版本一致，需要用/usr/lib/hbase/lib/目录下得jar包做软连接，如下图：

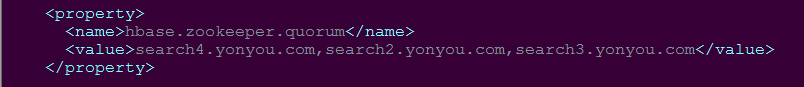


2.在UDH hive配置界面以下属性：

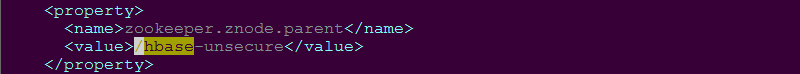
* hive.aux.jars.path



* hbase.zookeeper.quorum

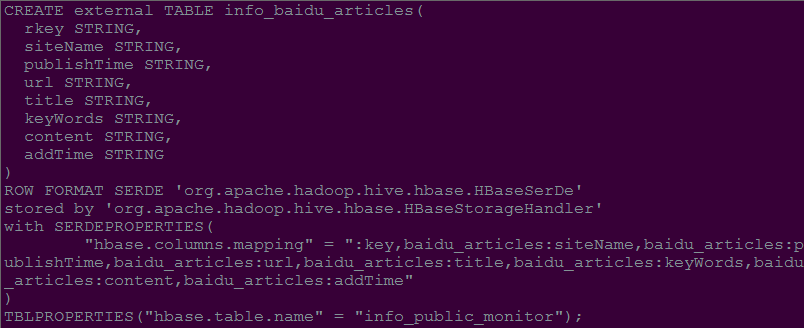


* zookeeper.znode.parent



### 二、Hive 映射已存在的Hbase表

Hive中的外部表可以映射到已存在的Hbase表，如映射到百度爬虫数据



上面DDL语句中，在WITH SERDEPROPERTIES选项中指定Hive外部表字段到HBase列的映射，其中“:key”对应于HBase中的RowKey，名称为“rkey”，其余的就是列簇info中的列名。最后在TBLPROPERTIES中指定了HBase中要进行映射的表名。

### 三、使用impala验证映射的Hbase表

Impala共享Hive的Metastore，这时需要同步元数据，可以通过在Hue中impala查询界面执行同步命令：

INVALIDATE METADATA;

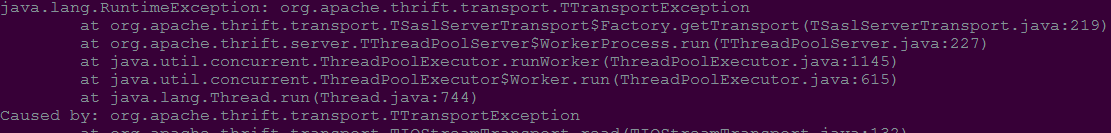
然后，就可以查看到映射HBase中表的结构：

* 
* 

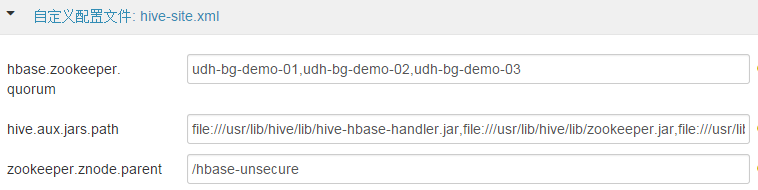
通过上面三步，我们就完成了Hive和HBase的集成配置

### 四、常见问题汇总

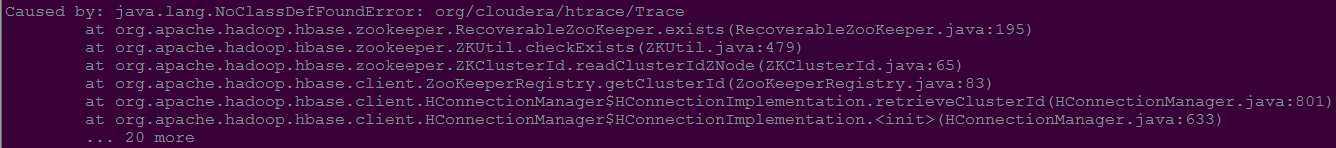
1. 在hive中创建映射表报如下错误，原因hive没有指定hbase的连接地址



解决：在hive-site.xml中添加Hbase的连接方法，需要添加两个属性分别是hbase.zookeeper.quorum、zookeeper.znode.parent对应如下图



1. 在hive中执行select count(\*) from scores语句查询时出现如下错误



解决：hive-site.xml中hive.aux.jars.path 属性里缺少 htrace-core.jar 库路径

1. Impala 执行查询出现AnalysisException: Failed to load metadata for table 错误，检查impala 配置主目录下（/etc/impala/conf）是否包含hbase-site.xml文件。