1. HBase整合Hive

1.1. 原理

Hive 与 HBase 利用两者本身对外的 API 来实现整合，主要是靠 HBaseStorageHandler 进行通信，利用 **HBaseStorageHandler**，Hive 可以获取到Hive表对应的 HBase 表名，列簇以及列，InputFormat 和 OutputFormat 类，创建和删除 HBase 表等。

Hive 访问 HBase 中表数据，实质上是通过 MapReduce 读取 HBase 表数据，其实现是在 MapReduce 中，使用HiveHBaseTableInputFormat 完成对 HBase 表的切分，获取 RecordReader 对象来读取数据。

对 HBase 表的切分原则是一个 Region 切分成一个 Split，即表中有多少个 Regions，MapReduce 中就有多少个MapTask。

读取 HBase 表数据都是通过构建 Scanner，对表进行全表扫描，如果有过滤条件，则转化为 Filter。当过滤条件为 RowKey 时，则转化为对 RowKey 的过滤，Scanner 通过 RPC 调用 RegionServer 的 next() 来获取数据。

```
HDFS：mapreduce执行计算的时候，默认是一个数据块执行一个MapTask
HBase: mapreduce执行计算的时候，默认是一个 region 执行一个task
```

1.2. 准备HBase表和数据

创建HBase表：

```
create 'mingxing', {NAME => 'base_info',VERSIONS => 1},{NAME =>
'extra_info',VERSIONS => 1}
```

插入准备数据：

```
put 'mingxing','rk001','base_info:name','huangbo'
put 'mingxing','rk001','base_info:age','33'
put 'mingxing','rk001','extra_info:math','44'
put 'mingxing','rk001','extra_info:province','beijing'
put 'mingxing','rk002','base_info:name','xuzheng'
put 'mingxing','rk002','base_info:age','44'
put 'mingxing','rk003','base_info:name','wangbaoqiang'
put 'mingxing','rk003','base_info:age','55'
put 'mingxing','rk003','base_info:gender','male'
put 'mingxing','rk004','extra_info:math','33'
put 'mingxing','rk004','extra_info:province','tianjin'
put 'mingxing','rk004','extra_info:children','3'
put 'mingxing','rk005','base_info:name','liutao'
put 'mingxing','rk006','extra_info:name','liujialing'
```

1.3. Hive端操作

进入Hive客户端，需要进行一下参数设置：

```
指定hbase所使用的zookeeper集群的地址：默认端口是2181，可以不写
set hbase.zookeeper.quorum=bigdata02:2181,bigdata03:2181,bigdata04:2181;

指定hbase在zookeeper中使用的根目录
set zookeeper.znode.parent=/hbase;

加入指定的处理jar
add jar /home/bigdata/apps/apache-hive-3.1.2-bin/lib/hive-hbase-handler-
3.1.2.jar;
```

创建基于HBase表的hive表：

所有列簇：

```
create external table mingxing(rowkey string, base_info map<string, string>,
extra_info map<string, string>)
row format delimited fields terminated by '\t'
stored by 'org.apache.hadoop.hive.hbase.HBaseStorageHandler'
with serdeproperties ("hbase.columns.mapping" = ":key,base_info:,extra_info:")
tblproperties("hbase.table.name"="mingxing","hbase.mapred.output.outputtable"="m
ingxing");
```

部分列簇部分列：

```
create external table mingxing1(rowkey string, name string, province string)
row format delimited fields terminated by '\t'
stored by 'org.apache.hadoop.hive.hbase.HBaseStorageHandler'
with serdeproperties ("hbase.columns.mapping" =
":key,base_info:name,extra_info:province")
tblproperties("hbase.table.name"="mingxing","hbase.mapred.output.outputtable"="m
ingxing");
```

部分参数解释：

```
org.apache.hadoop.hive.hbase.HBaseStorageHandler：处理hive到hbase转换关系的处理器
hbase.columns.mapping：定义hbase的列簇和列到hive的映射关系
hbase.table.name：hbase表名
```

1.4. 验证

查询语句：

```
select * from mingxing;
select count(*) from mingxing;
select count(rowkey) as total from mingxing;
select count(base_info['name']) as total from mingxing;
select rowkey,base_info['name'] from mingxing;
select rowkey,extra_info['province'] from mingxing;
select rowkey,base_info['name'], extra_info['province'] from mingxing;
```