大数据查询——HBase读写设计与实践

[http://mp.weixin.qq.com/s?\_\_biz=MzU1NDA4NjU2MA==&mid=2247486885&idx=1&sn=189abf32b7e8cb6ed0e9a26e8e7f5b21&chksm=fbe9b46acc9e3d7c02bbe4c34d862b86f3e347599bc1cdba65890aa02c31f742f022af3b3ee5&scene=27#wechat\_redirect](http://mp.weixin.qq.com/s?__biz=MzU1NDA4NjU2MA==&mid=2247486885&idx=1&sn=189abf32b7e8cb6ed0e9a26e8e7f5b21&chksm=fbe9b46acc9e3d7c02bbe4c34d862b86f3e347599bc1cdba65890aa02c31f742f022af3b3ee5&scene=27%23wechat_redirect)

读取Hbase可以使用基于Hbase的java API（spring-boot(Rest API)）

该链接分享中使用的是salt +hash散列+固定长度ID的方式来设计rowkey，可以保证rowkey的唯一性，以上设计能保证对每次查询而言，其 salt+hash 前缀值是确定的，并且落在同一个 region 中

Note:

当采用 bulk load 方式将HDFS中的数据写入多个预分区（region）时，要求 Spark 各 partition 的数据是有序的，RowKey，column family（cf），col name 均需要有序。

如果不进行二次排序会出现以下异常：(非字典排序的错误)



Spark计算能力很强，但是Hbase是基于LSM树的在提高了写的性能方面读的性能大大折扣，在HBASE表很大的时候，去读Hbase特别是全部扫描的情况下速度很慢。

Spark做离线分析

HIVE与HBASE的整合

使用Phoenix可以查询hbase数量小的表

Redis做在线数据查询，超出redis后查询落在hbase上

Apache kylin支持ad-hoc查询 SQL on Hbase OLAP分析