hugeGraph-Tools备份数据丢失实验报告

1、实验背景

我们公司在业务环境中需要定期用hugeGraph-Tools对backend为scylla的hugegraph进行备份，每次备份完成都会发现有大概千分之2的数据丢失。

2、实验基础配置环境

jdk：1.8

hugegraph：0.11.2

backend：scylla 4.3.2 standlone

ps：所有服务都是跑在windows10的笔记本电脑上，scylla运行在Oracle VM VirtualBox虚拟的centos7机器上，hugegraph运行在windows本地的idea中。

3、实验数据

- Schema

  VertexLabel：corporation

  IdStrategy：AUTOMATIC

Property	Datatype
citation	Integer
count	Integer
desc	String
descZh	String
hIndex	Integer
name	String
nameShortZh	String

nameZh	String
nameZhFirstChar	String

- 节点数据
  corporation.zip (json格式)

4、实验报告

用hugeGraph-Tools对这些测试数据进行back操作，实际数据量是1099047，下面是10次操作数据量的记录值

./hugegraph --url http://192.168.56.108:8080 backup -d ./

次数	数据量
1	1098074
2	1098395
3	1094354
4	1098063
5	1097578
6	1097258
7	1098148
8	1097755
9	1096583
10	1096804

5、复现过程和问题定位

通过跟踪hugeGraph-Tools备份节点数据的流程，发现主要调用了两个接口分别是graphs/{graph}/traversers/vertices/shards和graphs/{graph}/traversers/vertices/scan，下面就在本地idea手动调试这两个接口。

- 首先调用shards接口得到所有的shard：
  curl -X GET 'http://192.168.56.108:8080/graphs/hugegraph/traversers/vertices/shards?graph=hugegraph&split_size=67108864'

- 其中跳出一组shard作为scan的参数，来测试scan接口

  curl -X GET 'http://192.168.56.108:8080/graphs/hugegraph/traversers/vertices/scan?start=-2288559772019449 17&end=-150006368421838670&page=&page_limit=100000'

  - 多次用上面入参调用这个接口后，发现大多数情况下该接口返回4695条数据，偶尔返回少于4695条数据。
  - 从scan接口作为入口，经debug后发现当数据量少于4695条时，此时com.baidu.hugegraph.backend.page.PageEntryIterator类中的fetch方法少执行了一次，就此结果进一步观察，发现在此方法中的this.pageResults = this.queries.fetchNext(this.pageInfo, pageSize);这行代码在对scylla获取下一页数据的时候没有获取到。