|  |  |
| --- | --- |
| 交底书名称 | 一种校验hive与spark-sql的查询结果集的工具 |
| 技术联系人姓名 | 马殿军 |
| 技术联系人电话 | 13269325833 |
| 技术联系人Email | madianjun@jd.com |

（技术联系人信息用于与外部代理沟通，发明人信息在ERP专利申请系统中填写）

注意事项：

1、代理人并不是技术专家，交底书要使代理人能看懂，尤其是完整技术方案，一定要写得全面、清楚。

2、在后续与专利代理人进行沟通时，对于代理人的疑问应认真讲解，要求补充的材料应及时补充（禁止通过私人邮箱与代理人沟通）。

3、常用检索网站：www.soopat.com（SOOPAT），http://so.baiten.cn/（佰腾），patents.google.com（谷歌专利）。

# 1. 现有技术

/\* 应记载某个应用场景或者解决某个技术问题当前所采用的技术，可以概述该技术，也可以仅给出参考文献的链接或相关专利号。

当把sql语句从hive引擎迁移到spark sql引擎时，迁移之前要对语句的执行结果作校验；如果在两个引擎上跑的结果是一致的，则可以迁移，否则就不能迁移。当sql语句数量很庞大时，数据校验成本非常高，将给开发和运维工作带来巨大负担。本文针对该场景提出了一种用于迁移的校验工具，能够在迁移之前对比同一个语句在两个引擎的执行结果，避免迁移之后出现数据不一致的问题导致生产任务出现事故；同时也可提高对巨大sql语句数量的校验效率。

# 2. 现有技术的缺点

/\* 需要指出现有技术存在的缺点，本发明也不能克服的缺点无需提供。

/\* 应根据现有技术的实现过程，有针对性地说明缺点产生的原因。

目前没有现成的技术可用于迁移时的校验工作，大多依赖任务迁移之后用户自己发现故障，再经过开发人员排查问题原因，效率低下，可靠性很差。

# 3. 本发明技术方案

3.1 本发明所要解决的技术问题（即发明目的）

/\* 描述本发明所要解决的技术问题，与“2.现有技术的缺点”部分指出的缺点相对应。

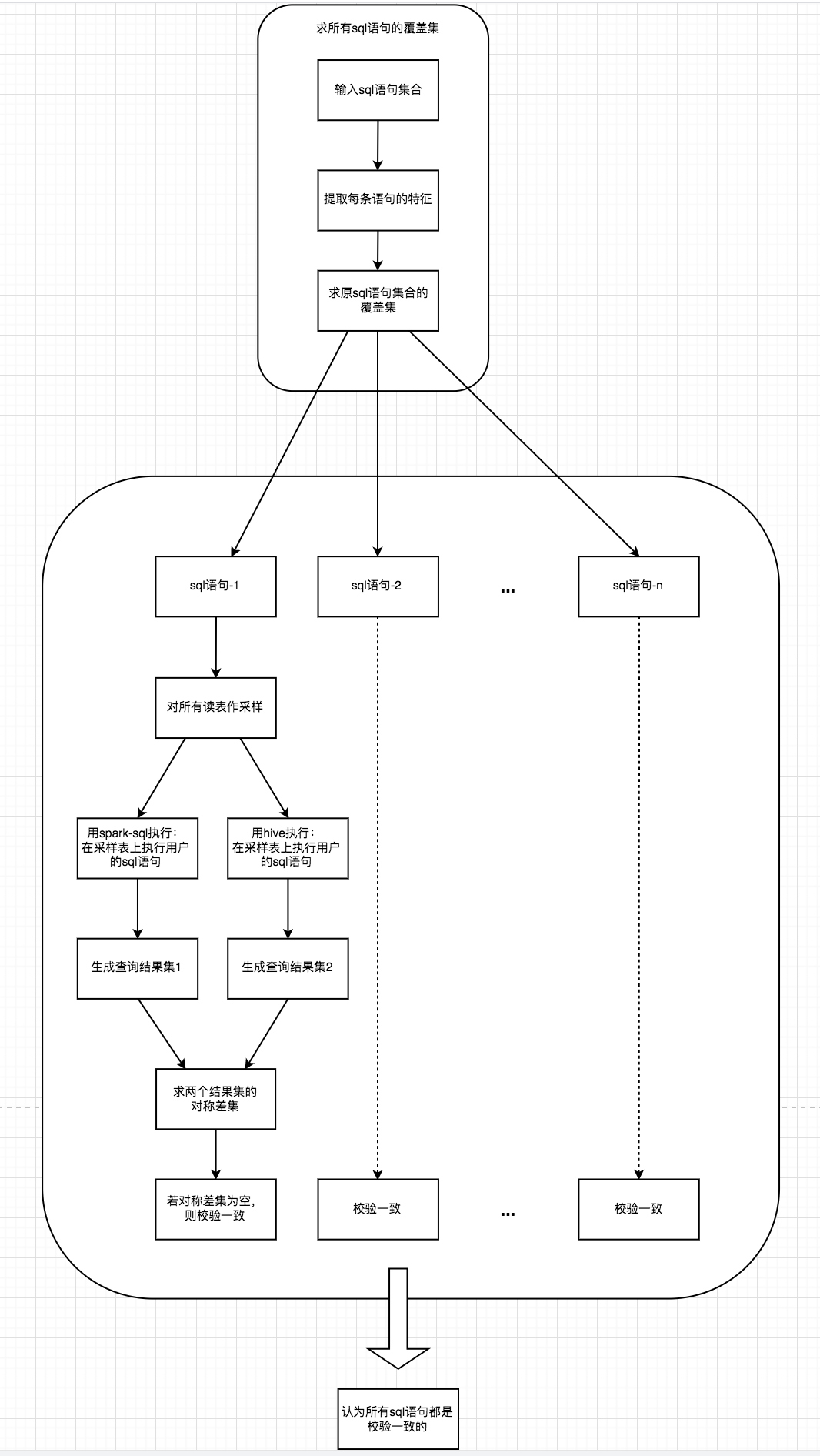
本发明主要解决的是，对大量sql语句在hive平台和spark-sql平台上执行的结果作对比校验，提供了一个自动化的校验工具，一方面使操作流程自动化，另一方面降低了需要处理的sql语句的规模，这样可以大大提高校验工作的效率。

3.2 本发明的完整技术方案的详细阐述

/\* 这是本文档最重要的部分，需要详细完整的阐述，不能光有原理，也不能仅有功能性介绍或操作说明。

/\* 在描述具体的技术方案时，必须结合附图（方法型专利按照数据流向或实现步骤抽象框图，装置型专利按照组成部件抽象框图）进行说明，每个附图都应当有对应的文字描述。如果本发明方案包含多个主题，方法与装置等，则需要分别进行描述。

整体流程图如下：



1. 求sql语句的覆盖集

将原始的大量的sql语句作为一个集合，利用“求覆盖集”的算法可得到一个较小的集合，该小集合可以表示原始语句集合的特征，这个小集合称为“覆盖集”。后续只需要对该覆盖集中的语句作校验即可，避免了处理原始的大量sql语句。在求覆盖集时，可以将sql语句的表名、算子、表达式等作为特征进行提取，特征选取得越合理、越全面，则覆盖集就能越充分地表示原始语句集合。

1. 校验覆盖集中的sql语句

针对覆盖集中的每条sql语句，做以下处理：

1. 重写sql语句：我们需要校验的是查询操作产生的结果集，只有以下三种操作跟“查询”操作有关（非查询操作不产生结果集，也就没有必要校验）：

(a)原语句是select操作，则在外层套上“create table <临时表> as” ，该“临时表”用来存储结果集。

(b)原语句是insert操作，即“insert into/overwrite table <表1> select … from <表2>”，则“表1”用来存储结果集。

(c)原语句是create操作，即“create table <表1> as select … from <表2>”,则“表1”用来存储结果集。

以上用来存储结果集的表，应当在实现时作重命名，以方便以下第(4)步中求对称差集。

1. 对读表作采样：所谓“读表”就是该sql语句需要查询的表，即“select … from …”所操作的表。有些sql语句执行时间较长，为了缩短校验时执行语句的时间，对读表作采样就能使数据量减小；如果sql语句执行时间本身就很短，也可以不采样。采样可以使用hive的“tablesample”操作。
2. 基于采样表执行sql语句：基于采样表的数据，将重写后的sql语句分别在hive和spark-sql上执行，分别得到各自的结果集。
3. 对比结果集：将第(2)步的两个结果集作对比校验，方法是使用sql语法中的“差集”操作（即except算子）得到对称差集，然后判断对称差集是否为空。假设两个结果集分别为A和B，则对称差集为。若对称差集为空，则说明两个结果集是完全相同的，那么校验是一致的；否则校验不一致。

当覆盖集中的sql语句都校验完成时，如果它们的校验结果是一致的，则认为原始的sql语句也是校验一致的；如果覆盖集中有校验不一致的语句，那么通过修改覆盖集算法，将校验不一致语句所表示的原始语句剔除（这样的原始语句也是不一致的），认为剩下的原始语句是一致的。

3.3 本发明希望保护的技术创新点

/\* 指出技术方案中希望保护的技术关键点，并概括说明该关键点的技术原理。

本发明的技术创新点是结合覆盖集算法、sql重写、采样、对称差集这些关键技术，实现一种对sql语句的查询结果集进行对比校验的工具，该工具使整个校验工作变得自动化，同时也大大减少了处理的sql语句数量，有利于提升校验工作的效率。

3.4 针对3.3中的技术方案，是否还有别的替代方案同样能完成发明目的？

/\* 替代方案可以是完整技术方案的替代，也可以是部分结构或者步骤的替代。

3.5交底书中技术术语的名词解释

/\* 记载交底书中出现的专业技术术语、缩写、外文的解释。

Spark：大规模数据处理计算引擎。

Spark SQL：是Spark中一个模块，用来处理结构化的数据。它的sql语法基本上与Hive兼容。

Hive：基于Hadoop的一个数据仓库工具，可以将结构化的数据文件映射为一张数据库表，并提供简单的sql查询功能，可以将sql语句转换为MapReduce任务进行运行。