Kettle 在分布式数据仓库领域的研究与实践

万峰华 余 臻 王 成 武萌 周铖辉

中国电子科技集团公司第二十八研究所 南京 210000

摘 要：针对业务系统多样化数据抽取加载需求，提出一种基于 Kettle 的分布式数据仓库构建实现方案。以 TPC-DS 工具构建模拟数据为基础，通过比较手动数据抽取、直接数据抽取和间接数据抽取三种方式的差异，选择间接数据抽取方法，配合操作系统定时任务，打破不同信息系统之间的“信息孤岛”，解决各业务系统间的互联互通和数据共享问题，完成源系统中的数据定期加载至数据仓库，实现分布式数据仓库的自动构建，为企业的分析和决策提供服务支撑。

关键词：Kettle；ETL；分布式数据仓库

0 引 言

随着移动互联网的飞速发展，信息化与工业化的快速融合，以指数增长的方式产生了海量的业务数据，如何从海量数据中抽取出有用的数据，供大数据分析和人工智能应用，成为当今的重要课题[1]。然而，大型企业、组织机构和政府部门由于业务信息系统之间的数据隔离和缺乏统一管理，不同厂商的系统互不兼容，不同系统之间无法互联互通和数据共享，形成一个个“信息孤岛”，使得各个系统之间的数据集成和共享变得困难。数据集成的目标是将来自不同来源、不同格式、不具备相同特性和属性的数据进行整合[2]，通过数据集成，用户可以更加高效地访问和利用数据，避免重复的数据采集和处理，同时提高数据分析的准确性和可信度。然而，传统的数据集成技术存在复杂度高、易错性高、性能低和不易扩展等问题。因此，通过设计一种基于 Kettle 的分布式数据仓库构建实现方案，高效地完成数据集成和转换任务，支撑企业的业务运作和决策分析。

1.Kettle 概述

Kettle 是一款免费开源、可视化、简单易用并且功能强大的 ETL 工具。Kettle 是使用纯 Java 编写的，运行在 Windows、Linux 和 Unix 之上，数据抽取简单并且高效，能满足各种场景的需求[3]。ETL 是数据仓库获取高质量数据的关键环节，是对分散在各业务系统中的现有数据进行抽取、转换、清洗和加载的过程，使这些数据成为商业智能系统需要的有用数据[4]。Kettle 提供了一个基于 Java 的可视化开发环境，可以通过拖拽组件并在组件之间配置数据流来创建 ETL 解决方案。ETL 活动是一个四元组 A=（ID，I，O，S），ID 是活动标示符，I 是输入模式的集合，O 是输出模式的集合，S 是一个或多个扩展的关系代数表达式，表示每个输出模式的语义[5-6]。KETTLE 的 ETL 活动可视为一个有向无环图（DAG 图），图的节点对应于一个个作业或转换步骤（Step），边代表数据供给关系对应于数据流节点连接（Hop），Kettle 的概念模型如下图 1 所示。

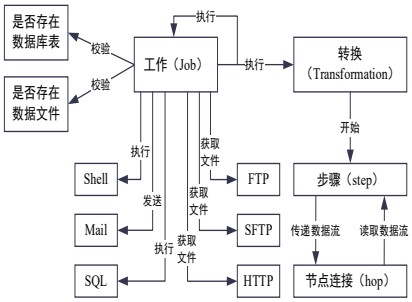

图 1 Kettle 概念模型图

2.TPC-DS 概述

TPC-DS（Transactional Processing Performance Council Decision Support Benchmark）是大数据领域常用的一种数据仓库基准测试套件，主要用于衡量数据仓库系统的性能。它模拟一个零售企业的决策支持环境，包括复杂的查询、数据挖掘和业务智能等功能，TPC-DS 标准在评估生产系统性能时，考虑了数据规模、查询复杂度、并发用户数等。TPC-DS 基准定义了 7 张事实表，17 张维度表，平均每张维度表含有 18 列，其采用了 99 个复杂的 SQL 查询案例，遵循

SQL99 和 SQL2003 标准，SQL 数量和难度上大大增加[7-8]。基准抽象出真实决策支持系统可能面临的多样性操作，同时保留了必要的性能特征，既可为企业当前的短期市场做决策，又可以预测企业将来可能面临的问题，为企业提供建议[9]。

TPC-DS 工具测试数据生成流程如下：

（1）工具下载。从官网上直接下载需要版本的 TPC-DS 源码。

（2）编译赋权。将源码包上传至服务器并解压，使用操作系统命令赋权，再进去源码包的 tools 目录下进行编译。如果编译报错，则需要手动挂载软件源并安装依赖包，最后再重试编译过程。

（3）DDL 生成。编译完成后，将在 tools 目录下生成包括 24 张表的 DDL 语句。

（4）测试数据生成。切换至 tools，设置数据规模参数和并发参数，执行 dsdgen 可执行程序，生成相应规模的数据集。

（5）模拟数据构建。在测试数据库 MariaDB 中，使用 DDL 语句创建表，再将生成的数据导入数据库中，模拟真实的业务数据存储在关系型数据库中的场景。

3. 数据抽取加载

3.1 手动数据抽取

以抽取存储在 MariaDB 数据库中的 TPC-DS 数据集到分布式数据仓库中为例。使用手动数据抽取的方式完成分布式数据仓库构建。以 call_center 表的为例，具体方式如下：

（1）数据导出。使用 MariaDB 的导出命令将数据导出至本地。

```
MariaDB [tpcds]> select * from tpcds.call_center into outfile '/opt/mariadb_output/call_center.csv';
```

（2）文件上传至分布式文件系统。

```
[hdfs@host76 ~]$ hadoop fs -mkdir /tmp/tpcds/
[hdfs@host76 ~]$ hadoop fs -chmod -R 755 /tmp/tpcds/
[hdfs@host76 ~]$ hadoop fs -put /opt/mariadb_output/call_center.csv /tmp/tpcds/
```

（3）hive 中创建表。参考 TPC-DS 工具 tools 下生成的 tpcds.sql 中 DDL 语句完成表的创建。

（4）导入数据。登录 hive，并进入指定数据库，执行导入语句。

```
[hdfs@host76 root]$ beeline
0: jdbc:hive2://host76:2181/default> use tpcds;
0: jdbc:hive2://host76:2181/tpcds> load data inpath '/tmp/tpcds/call_center.csv' into table call_center;
```

通过上述方法步骤可以将关系型数据库中的数据抽取至分布式数据仓库中，但是该方式全程手动操作，无法满足自动化数据抽取的应用场景，尤其当业务系统中的数据随时变化，可能需要实现数据的定期抽取。

3.2 直接数据抽取

考虑采用手动数据抽取方式无法适应分布式数据仓库构建的变化需求，利用 Kettle 工具，采用直接数据抽取方式完成分布式数据仓库构建。MariaDB 到 Hive 的数据抽取如下图 2 所示。以 call_center 表为例，具体方式如下：

（1）数据抽取流程构建

打开 Kettle 工具，新建转换，依次拖拽表输入和表输出至画布中并连接，初步完成数据抽取流程构建。

（2）关系型数据库设置

双击表输入算子，填写步骤名称信息，再点击新建，在弹出的数据库连接页面中选择 MariaDB 连接类型并填写连接参数，最后点击测试，测试通过后点击确认完成数据库连接创建。返回至表输入算子配置页面后，点击获取 SQL 查询语句，选择 tpcds 库和 call_center 表，完成查询 SQL 自动生成，最后点击确认完成表输入算子配置。

（3）分布式数据仓库设置

双击表输出算子，填写步骤名称信息，再点击新建，在弹出的数据库连接页面中选择 Hadoop Hive 2 连接类型并填写连接参数，最后点击测试，测试通过后点击确认完成分布式数据仓库连接创建。返回至表输出算子配置页面后，选择目标模式 tpcds 库和目标表 call_center，勾选数据库字段，点击获取字段，

完成字段流映射配置，最后点击确认完成表输出算子配置。如果分布式数据仓库中没有预先建立表，那么可以通过点击 SQL，再点击执行，完成表在目的端自动创建。

（4）数据抽取流程运行

按照上述步骤 1-3，完成 24 张表的数据抽取流程配置，最后点击保存并运行，完成 MariaDB 到 Hive 的数据抽取。

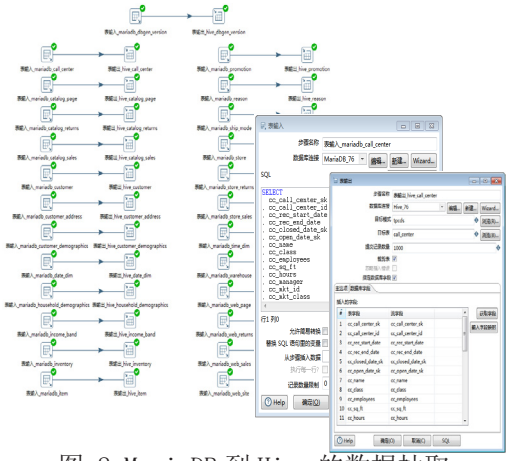

图 2 MariaDB 到 Hive 的数据抽取

通过上述方法步骤可以将关系型数据库中的数据抽取至分布式数据仓库中，尽管该方式提供可视化的配置界面，方便简洁易于操作，但是该方式的数据抽取流程运行耗时长，抽取速率慢，无法满足大规模数据集的应用场景。

3.3 间接数据抽取

考虑采用手动数据抽取方式或直接数据抽取方式进行分布式数据仓库构建所面临的问题，提出一种间接的数据抽取方式，既不需要过多的人工干预，数据抽取效率上又不会过慢，能够满足分布式数据仓库构建的变化需求。利用 Kettle 工具，采用直接数据抽取方式完成 MariaDB 到 HDFS 的数据迁移，再在分布式数据仓库上创建外表，并将表的 location 指向 HDFS 存储路径，完成数据文件与表的关联，实现分布式数据仓库构建。MariaDB 到 HDFS 的数据迁移如下图 3 所示。以 call_center 表为例，具体方式如下：

（1）数据抽取流程构建

打开 Kettle 工具，新建转换，依次拖拽表输入和 Hadoop file output 至画布中并连接，初步完成数据迁移流程构建。

（2）关系型数据库设置

参照直接数据抽取的步骤 2 完成表输入算子配置。

（3）分布式文件系统设置

双击 Hadoop file output 算子，填写步骤名称信息，再点击新建，在弹出的 Hadoop Cluster 配置页面中填写 HDFS、JobTracker 和 Zookeeper 等参数，最后点击测试，测试通过后点击确认完成 Hadoop 连接创建。返回至 Hadoop file output 算子配置页面后，选择文件存储路径，然后在内容页签下填写对应的分隔符，在字段页签下点击获取字段，完成字段流映射配置，最后点击确认完成 Hadoop file output 算子配置。

（4）数据抽取流程运行

按照上述步骤 1-3，完成 24 张表的数据迁移流程配置，点击保存并运行，完成 MariaDB 到 HDFS 的数据迁移。

（5）数据绑定

上述步骤中，数据已迁移至 /user/Administraor/call_center 目录下，参考 TPC-DS 工具 tools 下生成的 tpcds.sql 中 DDL 语句，创建 hive 表并将存储路径指向该目录，具体方式如下：

0: jdbc:hive2://host76:2181/tpcds>create table catalog_sales (...,cs_item_sk, cs_order_number,…) primary key (cs_item_sk, cs_order_number) row format delimited fields terminated by '|' location 'hdfs://host76:9020/user/Administraor/call_center';

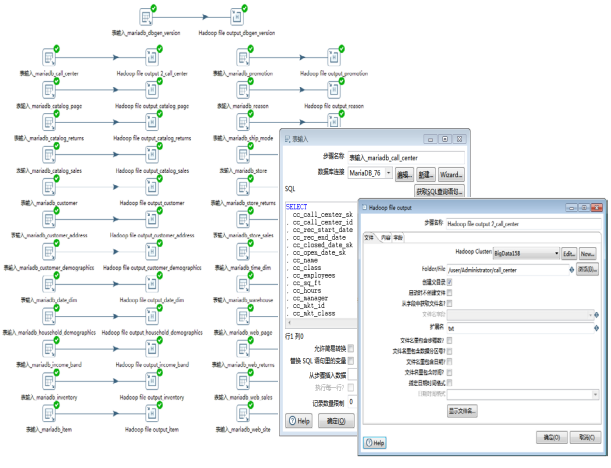

图 3 MariaDB 到 HDFS 数据迁移

3.4 数据抽取实施方案

考虑在实际情况下，基于间接数据抽取方式实现分布式数据仓库构建的实施方案，依然需要解决业务系统的数据会实时变化场景下的数据抽取。针对业务系统的数据会实时变化的数据抽取需求，可采用如下方法实现数据抽取转换任务的周期性执行。

（1）Kettle 工具的 Job 流程。打开 Kettle 工具，新建作业，依次拖拽 Start、转换和成功算子至画布中并连线，点击 Start 算子，可选择月、周、天、分钟和秒间隔的周期性配置，点击转换算子，浏览选择间接数据抽取的转换任务，完成数据作业流程构建。

（2）操作系统定时任务。考虑到 Kettle 工具在周期性执行方面的不便，通过编写自动化执行脚本，再由操作系统的定时任务进行调度，实现数据抽取转换任务的周期性执行。

如果 Kettle 工具部署在 Linux 机器上，使用操作系统的 crontab 调度脚本的周期性执行，其中 start.sh 脚本如下所示。

```
[root@host76 kettle]# cat start.sh
#!/bin/sh
/home/data-integration/kitchen.sh -file=/home/kettle/mysql_to_hdfs.ktr -level=Detailed -logfile=/home/kettle/mysql_to_hdfs.log
[root@host76 kettle]# crontab -l
*/1 * * * * sh /home/kettle/start.sh >> /home/kettle/mysql_to_hdfs_crontab.log
```

如果 Kettle 工具部署在 Windows 机器上，使用 Windows 操作系统自带的任务计划程序调度脚本的周期性执行，其中 start.bat 脚本如下所示。

```
D:\kettle>type start.bat
D:\data-integration\pan.bat /file D:\kettle\mysql_to_hdfs.ktr /level:Basic > D:\kettle\mysql_to_hdfs.log
```

4. 结束语

Kettle 在多源异构数据集成和数据抽取加载方面具有上手快、开发简单的优势，针对分布式数据仓库领域的应用需求，设计了基于 Kettle 的分布式数据仓库构建的实现方案，该方案依托 Kettle 和 TPC-DS 工具，在分析对比手动数据抽取、直接数据抽取和间接数据抽取差异的基础上，结合操作系统定时任务，实现了分布式数据仓库的自动构建，打破了业务系统之间的信息壁垒，可为企业的分析和决策提供服务支撑。未来将不断完善数据转换清洗的过程，提高数据质量，优化 ETL 设计流程，提升数据更新效率，为企业进一步的商业智能分析、OLAP 分析以及知识发现奠定数据基础。

参考文献：

[1] 王雪松，张良均．ETL 数据整合与处理 (Kettle)[M]. 北京：人民邮电出版社，2021,1-2.

[2] 崔记东．基于 Kettle 和 Quartz 的数据集成平台的研究与实现 [D]. 郑州大学，2019,12-13.

[3] 陈荣鑫，付永钢，陈维斌．基于 Pentaho 的商业智能系统 [J]. 计算机工程与设计 (9):263-265.

[4] 钟 华，冯文澜，谭红星，等．面向数据集成的 ETL 系统设计与实现 [J]．计算机科学，2004,31(9)：87-89.

[5] 吴远红．ETL 执行过程的优化研究 [J]. 计算机科学，2007,34(1):81-83.

[6] 崔有文，周金海．基于 Pentaho 的中药饮片企业商业智能研究 [J]. 电子设计工程 ,2014,22(7):12-15.

[7] 何磊．基于 TPC-DS 的测试系统研发 [D]. 上海．复旦大学 ,2013

[8] 牛一捷，邓武．决策支持评测系统的设计与实现 [J]. 计算机时代 ,2007(3):32-33

[9] 刘宝星．基于 TPC-DS 的性能测试工具设计与实现 [D]. 大连理工大学 ,2018.

作者简介：万峰华（1990-）男，汉，江西南昌，硕士研究生，中级工程师，研究方向：大数据处理。