**Adv\_DataX for Postgresql 架构说明**

**------天津第一DBA**

一 软件功能说明：

软件使用python开发，代码量200行以内，主要用作同构，异构数据库，数据表之间同步。

二 与阿里云DATAX相比：

|  |  |  |
| --- | --- | --- |
|  | 阿里云DATAX | Adv\_DataX for Postgresql |
| 数据库支持 | 大部分主流 | 目前仅支持postgresql |
| 性能 | 100分 | 101分？ |
| 资源占用 | 高内存，高cpu | 可以根据需要控制资源占用，包括cpu，iops |
| 代码复杂度 | Java，体积庞大 | Python，代码总量300以内 |
| 维护难以程度 | 需要填写excel，生成xml文件，表结构变动需要重新制作excel，不适用结构经常变动的业务 | 几行配置参数，不需要其他操作，其余信息自动从数据字典获取 |
| Upsert支持 | Postgresql中仅支持9.5以上版本 | Postgresql版本不限 |
| 是否借助中间表 | 是，增量数据需要写入临时表再写回中间表，需要预留足够的磁盘空间 | 不需要 |

三 架构说明：

软件在获取数据方面分为线程型和进程型。

进程型对CPU资源利用更好，速度更快，但是对主从两个节点会造成很大的压力。主库IOPS会满，从库的CPU会100%。但是仍然可以设置单一进程运行，此时同步顺序为按照预定的顺序对表进行同步。

线程型对硬件资源的压制更好，同步速度慢，对主从两个节点的压力非常小。可以根据预定的表同步顺序设定同时同步几张。例如该模式下，设定获取数据的线程为10，插入线程为20，运行时一次对10张表进行同步，每个同步任务再衍生出20个线程进行插入。当有一张表先同步完毕，系统会从队列中获取下一个需要同步的表。

同步速度主要受插入速度影响，虽然仍有提高的空间，但本人选择了相对稳妥的方式。



四： 特色功能

1 性能优秀，硬件资源影响程度可控

2 配置简单灵活

3 增量同步全系列数据库支持upsert功能，包括pg9.5之前的系统。但增量同步速度比全量同步稍慢。

五 适用范围

可以用于云端数据库之间同步，包括但不限于冷热分离。配置简单，同步规则可以通过其他脚本进行批量下发。

六 配置文件展

#同步表清单

MoveTableList=['table1','table2','table3','table4','table5']

#数据获取方式,可以根据具体的场景编写不同的SQL

DataFetchRuleSql="select \* from %s where xxx > xxx and xxx<xxx "

#最大线程/进程数 （Thread代表线程 Process 代表进程 本人此处偷懒，均使用Thread表示）

SourceMaxThread=4

#最大插入线程

InsertMaxThread=20

#每张表每次获取结果集的数据量

ResultCount=2000

#源端参数

sourceuser="user1"

sourcepassword="1234"

sourcehost="pgm-xxx.pg.rds.aliyuncs.com"

sourceport=5431

#目标端连接参数

sourcedbname="dbname"

destuser="user2"

destpassword="1234"

desthost="pgm-yyy.pg.rds.aliyuncs.com"

destport=5432

destdbname="test"