

云享家 | 数据库同步的正确打开方式

张小娅 BespinGlobal 1周前



前言

相信很多的用户在迁移上云上最头疼的是数据库备份和还原了吧。对于有些年头企业的数据可能很大，用数据库自带的dump工具备份导出又导入，在传输上耗时又费力。

BespIn Global资深交付工程师张小娅将为您介绍一种离线数据同步工具DATAX，实现包括MySQL、Oracle、SqlServer、Postgre、HDFS、Hive、ADS、HBase、TableStore(OTS)、MaxCompute(ODPS)、DRDS 等各种同构/异构数据源之间高效的数据同步功能。

教程

System Requirements

- Linux
- JDK(1.8以上，推荐1.8)
- Python(推荐Python2.6.X)
- Apache Maven 3.x (Compile DataX)

工具部署

- 直接下载datax工具包

<http://datax-opensource.oss-cn-hangzhou.aliyuncs.com/datax.tar.gz>

下载后解压至本地某个目录，进入bin目录，即可运行同步作业

```
$ cd {YOUR_DATA_HOME}/bin
$ python datax.py {YOUR_JOB.json}
```

• 配置示例

可以通过命令查看配置模板： `python datax.py -r {YOUR_READER} -w {YOUR_WRITER}`

• 重要参数配置说明

splitPk	MysqlReader进行数据抽取时，如果指定splitPk，表示用户希望使用splitPk代表的字段进行数据分片，DataX因此会启动并发任务进行数据同步，这样可以大大提供数据同步的效能。	否	空
Where	筛选条件，MysqlReader根据指定的column、table、where条件拼接SQL，并根据这个SQL进行数据抽取。	否	空
Querysql	在有些业务场景下，where这一配置项不足以描述所筛选的条件，用户可以通过该配置型来自定义筛选SQL	否	无
Batchsize	一次性批量提交的记录数大小，该值可以极大减少DataX与Mysql的网络交互次数，并提升整体吞吐量。但是该值设置过大可能会造成DataX运行进程OOM情况。	否	1024
Channel	传输启动的通道数，通道数和batchsize对性能影响较大	否	无

功能最佳实战

这里以同构数据库MySQL迁移为例向大家介绍迁移中JSON文件的具体配置。

*由于本文章内容过多迁移成功后的输出日志本文章中不体现

1 | IDC 数据库上云最佳实战

- 自建Mysql服务迁移至阿里云RDS-Mysql

```
{
  "job": {
    "setting": {
      "speed": {
        "channel": 5
      }
    },
    "content": [{
      "reader": {
        "name": "mysqlreader",
        "parameter": {
          "username": "root",
          "password": "xxxxx",
          "column": [
            "*"
          ]
        }
      }
    ]
  }
}
```



```
{
  "job": {
    "setting": {
      "speed": {
        "channel": 5
      }
    },
    "content": [{
      "reader": {
        "name": "mysqlreader",
        "parameter": {
          "username": "member",
          "password": "xxxxx",
          "column": [
            "*"
          ],
          "connection": [{
            "table": [
              "member"
            ],
            "jdbcUrl": [
              "jdbc:mysql://bj-cdb-16u291gu.sql.tencentcdb.com:63061/member?useUnicode=true&characterEncoding=gbk"
            ]
          ]
        }
      },
      "writer": {
        "name": "mysqlwriter",
        "parameter": {
          "username": "member",
          "password": "xxxxxxx",
          "column": [
            "*"
          ],
          "connection": [{
            "table": [
              "jdbc:mysql://rm-2zedc22cukg7v0hvifo.mysql.rds.aliyuncs.com:3306/mysqltest?useUnicode=true&characterEncoding=gbk",
            ]
          ]
        }
      }
    ]
  }
}
```

```

        "member"
      ]
    }]
  }
}

```

阿里云RDS-mysql数据库迁移至腾讯云CDB-mysql数据库类型，只需要将jdbcurl地址修改正确就可以

3 | 同一张表分多进程同步

一张表数据量较大时，可以启动多个datax进程进行传输，通过表里的列进行数据分批，比如日期或者自增列的键值，json写法如下：

```

{
  "job": {
    "setting": {
      "speed": {
        "channel": 4
      }
    },
    "content": [{
      "reader": {
        "name": "mysqlreader",
        "parameter": {
          "username": "root",
          "password": "xxxx",
          "connection": [{
            "jdbcUrl": [
              "jdbc:mysql://10.10.10.1:3306/bespIn?useUnicode=true&characterEncoding=gbk"
            ],
            "querySql": [
              "SELECT id,name from table1 WHERE id>499999"
            ]
          }
        ]
      }
    },
    "writer": {
      "name": "mysqlwriter",
      "parameter": {

```

```
"writeMode": "insert",
"username": "root",
"password": "xxxxx",
"column": [
    "*"
],
"connection": [{
    "jdbcUrl": "jdbc:mysql://10.10.0.2:3306/bespin?useUnicode=true&
characterEncoding=gbk",
    "table": [
        "table1"
    ]
}]
}
}
}
}
```

结束语

DataX的性能和稳定性均经过众多项目的验证，可以作为一款值得依赖的云端/线下数据迁移工具；同时，DataX支持众多的数据源，经过良好规划，可以实现异构数据之间的互通，降低数据迁移成本、提高迁移便捷性；另外，DataX还可作为一款ETL工具，也为数据进出大数据平台提供传输通道。当然，工具是为人所用，丰富的经验积累才能充分发挥出工具的强大功能，**欢迎大家和 Bespin 交付团队共同探讨云端应用为企业带来的价值实现。**



往期回顾

云享家 | 没有绝对的安全，只有在阿里云的最佳实践。

云享家 | 医疗和金融行业在AWS平台的合规配置

云享家 | 基于AWS云平台的IDS联合监控分析



点击阅读原文查看更多

阅读原文