DataX

第一章 概述

什么是DataX

DataX是阿里巴巴开源的一个异构数据源`离线同步工具`，致力于实现包括关系型数据库（MySQL、Oracle等）、HDFS、Hive、ODPS、HBase、FTP等各种异构数据源之间稳定高效的数据同步功能。

DataX的设计

为了解决异构数据源同步问题，DataX将复杂的网状的同步链路变成了星型数据链路，DataX作为中间传输载体`负责连接各种数据源`。当需要接入一个新的数据源的时候，只需要将此数据源对接到DataX，便能跟已有的数据源做到无缝数据的同步。

支持的数据源

DataX目前已经有了比较全面的插件体系，主流的RDMBS数据库、NOSQL、大数据计算系统都已经接入。

第二章 使用案例

从stream流读取数据并打印到控制台

查看配置模板(可选)

```
python datax.py -r streamreader -w streamwriter
```

编写配置文件

```
vim stream2stream.json
```

```
{
  "job": {
    "content": [
      {
        "reader": {
          "name": "streamreader",
          "parameter": {
            "sliceRecordCount": 10,
            "column": [
              {
                "type": "long",
                "value": "10"
              },
              {
                "type": "string",
                "value": "hello，DataX"
              }
```

```
                ]
            }
        },
        "writer": {
            "name": "streamwriter",
            "parameter": {
                "encoding": "UTF-8",
                "print": true

            }
        }
    }
  ],
  "setting": {
    "speed": {
      "channel": 1
    }
  }
 }
}
```

运行

```
/opt/module/datax/bin/datax.py /opt/module/datax/job/stream2stream.json
```

读取MySQL中的数据存放到HDFS

创建student表

```
mysql> create database datax;
mysql> use datax;
mysql> create table student(id int,name varchar(20));
```

插入数据

```
mysql> insert into student values(1001,'zhangsan'),(1002,'lisi'),(1003,'wangwu');
```

编写配置文件

```
vim /opt/module/datax/job/mysql2hdfs.json
```

```
{
"job": {
"content": [
{
"reader": {
"name": "mysqlreader",
"parameter": {
"column": [
"id",
"name"
],
"connection": [
                        {
                                                "jdbcUrl": [
                                                        "jdbc:mysql://master:3306/datax"
                                                ],
                                                "table": [
```

```
                                                "student"
                                        ]
                                }
                        ],
                        "username": "root",
                        "password": "000000"
                }
        },
        "writer": {
                "name": "hdfswriter",
                "parameter": {
                        "column": [
                                {
                                        "name": "id",
                                        "type": "int"
                                },
                                {
                                        "name": "name",
                                        "type": "string"
                                }
                        ],
                        "defaultFS": "hdfs://master:9000",
                        "fieldDelimiter": "\t",
                        "fileName": "student.txt",
                        "fileType": "text",
                        "path": "/",
                        "writeMode": "append"
                }
        }
    }
  ],
  "setting": {
    "speed": {
  "channel": "1"
  }
  }
  }
  }
```

执行任务

```
bin/datax.py job/mysql2hdfs.json
```

查看HDFS

读取HDFS数据写入MySQL

```
vim job/hdfs2mysql.json
```

```
{
    "job": {
        "content": [
            {
                "reader": {
                    "name": "hdfsreader",
                    "parameter": {
                        "column": ["*"],
```

```
                                    "defaultFS": "hdfs://hadoop102:9000",
                                    "encoding": "UTF-8",
                                    "fieldDelimiter": "\t",
                                    "fileType": "text",
                                    "path": "/student.txt"
                              }
                        },
                        "writer": {
                              "name": "mysqlwriter",
                              "parameter": {
                                    "column": [
                                          "id",                                          "name"

                                    ],
                                    "connection": [
                                          {
                                                "jdbcUrl": "jdbc:mysql://hadoop102:3306/datax",
                                                "table": ["student2"]
                                          }
                                    ],
                                    "password": "000000",
                                    "username": "root",
                                    "writeMode": "insert"
                              }
                        }
                  }
            ],
            "setting": {
                  "speed": {
                        "channel": "1"
                  }
            }
      }
}
```