使用 DataX 导入或导出 COS

最近更新时间：2022-03-31 17:28:29

环境依赖

- HADOOP-COS 与对应版本的 cos_api-bundle。
- DataX 版本：DataX-3.0。

下载与安装

获取 HADOOP-COS

在官方 Github 上下载 HADOOP-COS 与对应版本的 cos_api-bundle。

获取 DataX 软件包

在官方 Github 上下载 DataX。

安装 HADOOP-COS

下载 HADOOP-COS 后，将 `hadoop-cos-2.x.x-${version}.jar` 和 `cos_api-bundle-${version}.jar` 拷贝到 Datax 解压路径 `plugin/reader/hdfsreader/libs/` 以及 `plugin/writer/hdfswriter/libs/` 下。

使用方法

DataX 配置

修改 datax.py 脚本

打开 DataX 解压目录下的 bin/datax.py 脚本，修改脚本中的 CLASS_PATH 变量为如下：

```
CLASS_PATH = ("%s/lib/*:%s/plugin/reader/hdfsreader/libs/*:%s/plugin/writer/hdfswriter/libs/*:.") % (DATAX_HOME, DATAX_HOME, DATAX_HOME)
```

在配置 JSON 文件里配置 hdfsreader 和 hdfswriter

示例 JSON 如下：

```
{
  "job": {
    "setting": {
      "speed": {
        "byte": 10485760
      },
      "errorLimit": {
        "record": 0,
        "percentage": 0.02
      }
    },
    "content": [{
      "reader": {
        "name": "hdfsreader",
        "parameter": {
          "path": "testfile",
          "defaultFS": "cosn://examplebucket-1250000000/",
          "column": ["*"],
          "fileType": "text",
          "encoding": "UTF-8",
          "hadoopConfig": {
            "fs.cosn.impl": "org.apache.hadoop.fs.CosFileSystem",
            "fs.cosn.userinfo.region": "ap-beijing",
            "fs.cosn.tmp.dir": "/tmp/hadoop_cos",
            "fs.cosn.userinfo.secretId": "COS_SECRETID",
            "fs.cosn.userinfo.secretKey": "COS_SECRETKEY"
          },
          "fieldDelimiter": ","
        }
```

```
        },
        "writer": {
          "name": "hdfswriter",
          "parameter": {
            "path": "/user/hadoop/",
            "fileName": "testfile1",
            "defaultFS": "cosn://examplebucket-1250000000/",
            "column": [{
                  "name": "col",
                  "type": "string"
                },
                {
                  "name": "col1",
                  "type": "string"
                },
                {
                  "name": "col2",
                  "type": "string"
                }
            ],
            "fileType": "text",
            "encoding": "UTF-8",
            "hadoopConfig": {
              "fs.cosn.impl": "org.apache.hadoop.fs.CosFileSystem",
              "fs.cosn.userinfo.region": "ap-beijing",
              "fs.cosn.tmp.dir": "/tmp/hadoop_cos",
              "fs.cosn.userinfo.secretId": "COS_SECRETID",
              "fs.cosn.userinfo.secretKey": "COS_SECRETKEY"
            },
            "fieldDelimiter": ":",
            "writeMode": "append"
          }
        }
      }]
    }
}
```

配置说明如下：

- hadoopConfig 配置为 cosn 所需要的配置。
- defaultFS 填写为 cosn 的路径，例如 `cosn://examplebucket-1250000000/` 。
- fs.cosn.userinfo.region 修改为存储桶所在的地域，例如 `ap-beijing` ，详情请参见 地域和访问域名 。
- COS_SECRETID 和 COS_SECRETKEY 修改为 COS 密钥。

其他配置同 hdfs 配置项即可。

执行数据迁移

将配置文件保存为 hdfs_job.json，存放到 job 目录下，执行以下命令行：

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

观察屏幕正常输出如下：

```
2020-03-09 16:49:59.543 [job-0] INFO  JobContainer -
        [total cpu info] =>
             averageCpu                     | maxDeltaCpu                    | minDeltaCpu
             -1.00%                         | -1.00%                         | -1.00%


        [total gc info] =>
             NAME                 | totalGCCount       | maxDeltaGCCount    | minDeltaGCCount    | totalGCTime        | maxDeltaGCTime     |
minDeltaGCTime
             PS MarkSweep         | 1                  | 1                  | 1                  | 0.024s             | 0.024s             | 0.024s
             PS Scavenge          | 1                  | 1                  | 1                  | 0.014s             | 0.014s             | 0.014s

2020-03-09 16:49:59.543 [job-0] INFO  JobContainer - PerfTrace not enable!
```

```
2020-03-09 16:49:59.543 [job-0] INFO  StandAloneJobContainerCommunicator - Total 2 records, 33 bytes | Speed 3B/s, 0 records/s | Error
0 records, 0 bytes |  All Task WaitWriterTime 0.000s |  All Task WaitReaderTime 0.033s | Percentage 100.00%
2020-03-09 16:49:59.544 [job-0] INFO  JobContainer -
任务启动时刻                    : 2020-03-09 16:49:48
任务结束时刻                    : 2020-03-09 16:49:59
任务总计耗时                    :                 11s
任务平均流量                    :               3B/s
记录写入速度                    :              0rec/s
读出记录总数                    :                   2
读写失败总数                    :                   0
```