# Dship功能介绍

## 安装

1. 正确安装java运行环境，jdk1.6以上
2. 下载dship安装文件，odps-dship.zip，解压缩
3. 运行dship config命令配置accessid，accesskey及默认的project，其它见帮助, ./dship help 。

## 功能简介

1. 上传：支持文件和目录（只一级目录）的上传，每一次上传只支持数据上传到一个table或一个partition。

例：dship upload log.txt test\_project.test\_table/p1="b1",p2="b2"

1. 下载：只支持下载到单个文件，每一次下载只支持下载一个table或一个partition到一个文件。

例：dship download test\_project.test\_table/p1="b1",p2="b2" log.txt

1. 预扫描数据：上传之前默认会预扫描本地文件，通过表schema对数据内容进行校验。

例(只做预扫描，不实际上传)：dship upload log.txt test\_table --scan=only

1. 续传：因为网络或tunnel服务的原因出错，dship支持文件或目录的续传。

例：dship rerume

1. 脏数据：如果设置discard-bad-records=true，脏数据文件会保存在session目录中。

例：dship show bad

1. 查看session历史：可以通过show命令查看session相关信息，用purge命令删除session，具体见help。

## 数据类型

### STRING：

字符串类型，长度不能超过2M.

### BOOLEN:

布尔类型说明:

1. 上传值只支持”true”，”false”，”0”，”1”
2. 下载值为true\false
3. 不区分大小写

### BIGINT：

长整型：

1. 最大值：9223372036854775807
2. 最小值: -9223372036854775807

### DOUBLE:

双精度浮点数：

1. 有效位数16位
2. 上传支持科学计数法表示
3. 下载仅使用数字表示
4. 最大值：1.7976931348623157E308
5. 最小值：4.9E-324
6. 无穷大：Infinity
7. 无穷小：-Infinity

### DATETIME：

可以通过命令行指定用户数据日期格式的format pattern。

如：dship upload log.txt test\_table –dfp “yyyyMMddHHmmss”

pattern参考：

1. “yyyyMMddHHmmss”（默认） : 数据格式”20140209101000”
2. “yyyy-MM-dd HH:mm:ss”：数据格式” 2014-02-09 10:10:00”
3. “yyyy年MM月dd日” : 数据格式” 2014年09月01日”

pattern可参考： SimpleDateFormat

<http://docs.oracle.com/javase/6/docs/api/java/text/SimpleDateFormat.html>

### DECIMAL: （暂时不对外）

小数点后有效位数10位

## 分隔符

dship支持用户自定义的文件分隔符，行分隔符为“--record-delimiter”，列分隔符为“--field-delimiter”。

例：dship upload log.txt test\_table –fd “||” –rd “\r\n”

分隔符说明：

1. 支持多个字符的行列分隔符
2. 列分隔符不能够包含行分隔符
3. 转意字符分隔符，在命令行方式下只支持\r、\n和\t

## 空值

所有数据类型都可以有空值。

1. 默认空字符串为空值。
2. 可在命令行下通过--null-indicator参数来指定空值的字符串

例：dship upload log.txt test\_table –ni “NULL”

## 字符编码

用户可以指定文件的字符编码，默认为”UTF-8”。

例：dship upload log.txt test\_table –c “gbk”

## 常见问题

1. Q:是否支持ascii字符的分隔符：

A:命令行方式不支持，配置文件可以用十六进制表示：如\u000A，表示回车

1. Q:文件大小是否有限制：

A:文件大小没有限制，但一个upload上传时间最多24小时，工具会根据当前上传文件的速度预测，如果上传时间预计超过20小时，会中止上传。如：当前上速度20MB/s，则最多一个上传文件或目录大小为1.4T

1. Q: 记录大小是否有限制:

A:字符串最大2MB(utf8)，最大不超过200M，原因，用户可能指定错列分隔符。

1. Q:是否要使用压缩：

A:默认会使用压缩，如果带宽允许的情况下，可以关掉压缩，单次upload上传可能会快一些

1. Q:同一个表或partition是否可以并行上传：

A:可以

1. Q:什么时间续传使用--force参数:

A:如果使用上传的进程被强制中断，可以使用—force参数强制上传。

1. Q:是否支持不同字符编码：

A:支持不同的编码格式参数，带bom标识文件不用指定编码。

1. Q:导入成功脏数据怎么处理：

A:导入成功，如果有脏数据可以通过./dship show dad [sessionid]导出脏数据，修改导出的脏数据文件再次上传，可以通过./dship show history来显示那些session有脏数据。