# 数据存储

数据存储是DMP平台专门负责数据读写的模块。对外提供一个抽象的读写接口。数据处理模块和数据应用模块的程序可能会调用数据存储中已有的数据，而这两个模块的输出结果部分也会写入数据存储模块中。

数据存储单独作为一个模块的目的是：1. 不同的数据应用共享同一数据源，避免同一数据源的冗余存储。 2. 更新数据源时，保证不同的应用中版本一致，例如知识库的统一管理。3. 便于和任务平台隔离，以免因为存储模块消耗内存过多而使任务终端或平台崩溃。4. 底层使用分布式存储，根据数据量大小灵活添加或移除机器，使数据存储具有良好的扩展性。5. 使开发人员能专注于开发高效的实时读写服务，为用户定向、广告匹配等提供服务。6. 便于提供一系列简单的调用接口，简化数据处理模块和数据应用模块程序的调用。

## 接口描述

数据存储的接口分为http接口和java接口两种。Http接口简单直接，便于跨语言调用。Java接口满足开发高效的数据应用。Http接口规范旨在指定请求url和响应的格式。Java 接口旨在指定该模块api。

* Http接口

提供如下REST风格的HTTP接口：

http://<ip>:<port>/<method>?{query\_string}

query\_string由系统级参数部分和具体Open API调用参数部分组成，以key1=value&key2=value2&…表示，对于采用POST请求的Open API，query\_string部分则是在POST请求体里。所有查询类的Open API接口既支持POST，也支持GET方式，提交类的OPEN API接口仅支持POST方式。

* 表格1 系统级参数

|  |  |  |  |
| --- | --- | --- | --- |
| **参数名** | **类型** | **是否必需** | **描述** |
| method | string | 是 | 采取userinfo.query这样的命名空间方式制定方法 |
| call\_id | int | 否 | 时间戳，系统时间的秒值,同个应用的不同api请求的time值应该是递增的, 用于防replay攻击 |
| format | string | 否 | 响应包格式，可以是xml（默认）或json |
| ie | string | 否 | API调用请求包的编码类型，支持UTF-8和GBK |
| ip | string | 是 | 提供存储查询服务的ip |
| port | int | 是 | 提供存储查询服务的端口 |

* 用户数据查询接口
* 参数

|  |  |  |  |
| --- | --- | --- | --- |
| **参数名** | **类型** | **是否必需** | **描述** |
| port | int | 是 | 6579 用户数据查询 |
| method | string | 是 | userinfo.query 查询用户特征 |
| uid | string | 是 | 用户标识，例如RADIUSID、COOKIE |
| feattype\* | int | 否 | 特征类型，0 百度搜索  1 百度搜索关键词  2 电商title  3 电商title关键词  4 电商搜索  5 电商搜索关键词 |
| starttime | long | 否 | 起始时间 |
| endtime | long | 否 | 结束时间 |

注：\*表示可以有0-n个

* 返回值
* Response XML 示例

<?xml version=*"1.0"* encoding=*"UTF-8"*?>

<clickwise\_userinfo\_response xmlns=[*http://clickwise.cn*](http://clickwise.cn) xmlns:xsi=*"http://www.w3.org/2001/XMLSchema-instance"*>

<user>

<uid>7d1cf3105a7388753da162a802d78055009</uid>

<infolist>

<feat>

<feattype>bdsearch</feattype>

<conent>植物大战僵尸小游戏</conent>

</feat>

<feat>

<feattype>bdkeywords</feattype>

<conent>植物大战僵尸 小游戏</conent>

</feat>

<feat>

<feattype>ectitle</feattype>

<conent>春秋平底平跟流苏单鞋唐卡豆豆鞋牛筋底磨砂真牛皮豹纹小白鞋女鞋 </conent>

</feat>

</infolist>

</user>

</clickwise\_userinfo\_response>

* CookieMap 接口
* 参数

|  |  |  |  |
| --- | --- | --- | --- |
| **参数名** | **类型** | **是否必需** | **描述** |
| port | int | 是 | 6578 cookie map |
| method | string | 是 | cookiemap.map 映射用户id |
| src\_uid | string | 是 | 源用户标识，例如RADIUSID、COOKIE |
| src\_type | int | 是 | 源用户标识类型  0 RADIUSID  1 COOKIE  2 Email |
| dest\_type+ | int | 是 | 目标用户标识类型  0 RADIUSID  1 COOKIE  2 Email |

注：+表示可以有1-n个

* 返回值
* Response XML 示例

<?xml version=*"1.0"* encoding=*"UTF-8"*?>

<clickwise\_cookiemap\_response xmlns=[*http://clickwise.cn*](http://clickwise.cn) xmlns:xsi=*"http://www.w3.org/2001/XMLSchema-instance"*>

<src\_uid>

<uid\_type>cookie</uid\_type>

<uid\_value>7d1cf3105a7388753da162a802d78055009</uid\_value>

</src\_uid>>

<dest\_uids>

<uid\_item>

<uid\_type>radius</uid\_type>

<uid\_value>854f8dddc446a9822572aa47f8f75500e</uid\_value>

</uid\_item>

<uid\_item>

<uid\_type>email</uid\_type>

<uid\_value>test@126.com</uid\_value>

</uid\_item>

</dest\_uids>

</clickwise\_cookiemap\_response>

* Java API

## 设计实现

数据存储共分为数据存储介质、数据存储驱动、存储查询服务三个子层次。



数据存储介质子层是实际存储数据的各种类型的数据库或文件系统。根据响应时间分为非实时存储介质，如文本文件、hdfs、hive、mysql等，和实时存储介质，如hbase、redis、ardb、memcache等。根据可扩展性分为本地存储介质，如文本文件、mysql、redis等，和分布式存储介质,如hdfs、hive、hbase等。根据所要存储的数据量大小、数据格式以及响应时间选择实际的存储介质。一些查询速度快的介质，如redis、内存等可以用来做本地缓存。

数据存储驱动子层是数据存储介质子层上面封装的一系列接口，该子层的存在目的是：1. 统一管理已有数据，能够方便的浏览已有的表、数据库以及各自的格式 2. 记录读写日志，以分析数据的质量情况、读写的速度和稳定性。3. 将高并发实时查询

## 安装部署