套餐推荐专题

后台脚本部署手册

|  |  |  |  |
| --- | --- | --- | --- |
| 编写人 | 史婷 | 编写时间 | 2015-12-21 |
| 审核人 |  | 审核时间 |  |
| 版本 | v1.0 | | |

**文档修订记录**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **序号** | **版本** | **修改内容说明** | **修改人** | **修改日期** | **审阅者** | **审阅日期** |
| 1 |  |  |  |  |  |  |
| 2 |  |  |  |  |  |  |
| 3 |  |  |  |  |  |  |

目录

[1 引言 1](#_Toc443488684)

[1.1 文档说明 1](#_Toc443488685)

[1.2 阅读对象 1](#_Toc443488686)

[1.3 参考资料 1](#_Toc443488687)

[2 硬件清单 1](#_Toc443488688)

[3 软件环境安装说明 2](#_Toc443488689)

[3.1 软件及服务搭建清单 2](#_Toc443488690)

[3.1.1 基础环境安装验证 2](#_Toc443488691)

[3.1.2 后台数据挖掘部署 3](#_Toc443488692)

[3.1.2.1 R语言安装 3](#_Toc443488693)

[3.1.2.2 挖掘机器模块安装 3](#_Toc443488694)

[4 后台程序部署 3](#_Toc443488695)

[4.1 程序部署步骤 3](#_Toc443488696)

[4.2 配置文件清单 4](#_Toc443488697)

[4.3 部署代码清单 5](#_Toc443488698)

[4.4 实施人员注意事项 6](#_Toc443488699)

[5 部署验证 6](#_Toc443488700)

[6 脚本调试参考 7](#_Toc443488701)

[7 数据核查办法 7](#_Toc443488702)

[8 常见问题处理 12](#_Toc443488703)

# 引言

## 文档说明

本文档对套餐推荐后台脚本部署流程做出说明，以便参与开发设计的人员和现场实施人员有进一步的了解，方便日后维护。本文档应尽量做到全面细致、正确无误，但限于编者水平有限，若其中有错误、不准确的地方敬请指出以便更正。

## 阅读对象

* 项目开发人员
* 系统交付维护人员
* 现场PSO人员

## 参考资料

|  |  |
| --- | --- |
|  |  |
| Svn | http://60.194.3.165:2350/svn/BI/minedata-app/doc/02需求类/河南流量二期主套餐推荐 |
| Svn | http://60.194.3.165:2350/svn/BI/minedata-app /doc/03设计类/05 详细设计/河南流量2期主套餐推荐 |
| Svn | http://60.194.3.165:2350/svn/BI/minedata-app /doc/03设计类/03 模型设计/07河南流量2期主套餐推荐 |

# 硬件清单

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 机器ip | 机器名称 | 用户 | 密码 | 平台 | | 安装说明 |
| 192.168.0.1 | YDSJ1 | ocdp | ocdp@123 | hadoop | Namenode | Mysql |
| 192.168.0.2 | YDSJ2 | ocdp | ocdp@123 | Namenode | Hive客户端 |
| 192.168.0.3~6 | YDSJ3~6 | ocdp | ocdp@123 | Datenode |  |
| 192.168.0.222 | LLYY1 |  |  |  | 前台库 | Db2/mysql/was |

说明，IP地址、用户名/密码、机器名称需要和现场确认

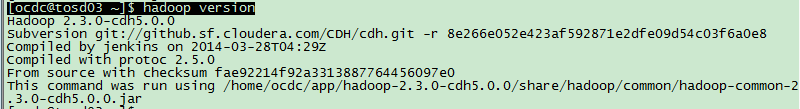
# 软件环境安装说明

## 软件及服务搭建清单

### 基础环境安装验证

1. Hadoop 2.3.0-cdh5.0.0

验证方法：[ocdc@tosd03 ~]$ hadoop version

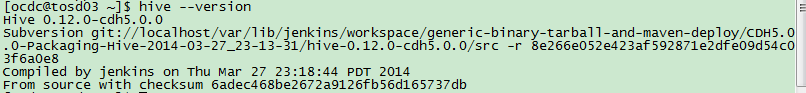


验证方法：[ocdc@tosd03 ~]$ jps



1. Hive 0.12.0-cdh5.0.0

验证方法：hive –version



1. Python 2.6.6

安装在hive所在节点上，安装方法参考《Python安装手册》

1. Thrift -0.9.2

安装在python的机器上，同上，所以安装在hive所在节点上，安装方法参考《Thrift安装手册》

1. ftp

安装参考第5章节部署验证 ，ftp验证部分。

### 后台数据挖掘部署

#### R语言安装

安装在所有hadoop节点机上，

安装方法参见《R语言安装手册》

#### 挖掘机器模块安装

安装在hive机器上，

安装方法参见《调度jar包安装部署文档.docx》

# 后台程序部署

## 程序部署步骤

后台脚本部署在hive客户端所在节点上,hive 端ip参见第2章节硬件清单

1. 创建目录

cd tas

mkdir setings（如果setings目录存在，就不用创建了）

mkdir plancmnd

1. 步骤二：获取脚本程序清单（脚本路径：svn路径minedata-app\trunk\backend-code\mda-recomm-mainplan）、（配置文件路径：203服务器，/home/ocdc/tas/setings）；
2. 步骤三：上传脚本到home/tas/plancmnd目录下，参见4.3

配置脚本上传到/home/ocdc/tas/setings

1. 步骤四：修改.pth文件，参见4.2小节；
2. 修改thrift.cfg文件，参见4.2小节；
3. 修改CommonThriftServer.py，参见4.2小节；
4. 修改init.cfg文件，参见4.2小节；
5. 部署验证，参见第五章节。
6. 数据源接入（历史用户信息表mda\_plan\_useraction\_yyyymm，B域用户数据信息表mda\_plan\_userdata\_yyyymm，套餐维表mda\_plan\_maincourse\_yyyymm，用户信息表mda\_plan\_userinfo\_yyyymm等表源数据接入，本地测试时可自己造数据txt格式，然后load到表中；现场依情况修改或增加数据源接入脚本）
7. 营销数据接入（效果评估需要营销数据平台数据给我们，目前没有接口，此部分预留，测试需要我们自己造数据测试mda\_plan\_market\_yyyymm）（目前实现方式是现场需要提供一个命名为 ：套餐id\_营销月份.txt 的文件到指定路径，供ftp下载解析）；
8. 带数据调试脚本

需要注意几个挖掘的脚本，参见第6章节

1. 数据核查，参见第7章
2. 配调度（根据本地使用的调度软件，咨询调度支撑人员）
3. 调度全流程测试

## 。配置文件清单

|  |  |  |
| --- | --- | --- |
| **配置文件** | **说明** | **修改说明** |
| hqltool.py | **Python脚本的参数配置文件，文件位置提交/home/ocdc/tas/setings目录** | **研发提供脚本** |
| settings.py | **Python脚本的参数配置文件，**  **文件位置提交/home/ocdc/tas/setings目录** | **研发提供脚本** |
| baseModule.py | **Python脚本的参数配置文件，文件位置提交/home/ocdc/tas/setings目录** | **研发提供脚本** |
| CommonThriftServer.py | **Python脚本的参数配置文件，文件位置提交/home/ocdc/tas/setings目录（用于前后台联调，启动thrift服务）** | **研发提供脚本**  **cd ./tas/setings**  **vi** CommonThriftServer.py  **修改transport = TSocket.TServerSocket('192.168.0.203',9090) 中ip地址为本地**IP地址  用于启动 主套餐推荐thrift服务 ，命令：python CommonThriftServer.py |
| thrift.cfg | **Python脚本的参数配置文件，文件位置提交/home/ocdc/tas/setings目录（用于前后台联调）** | **研发提供脚本**  **cd ./tas/setings**  **Vi t**hrift.cfg  **删除不相干的内容推荐服务接口部分，并修改主套餐推荐服务接口中的接口路径。** |
| .pth | **以上配置文件的存放路径，需要在python安装目录下的site-packages中（由于环境不同，请自行查找服务器上的该路径，例如：/usr/lib64/python2.6/site-packages）添加该文件并写入脚本文件路径。** | **cd usr/lib64/python2.6/site-package**  **[ocdc@tosd03 site-packages]$ vi .pth**  **/home/ocdc/tas/setings** |
| init.cfg | **Python脚本的参数配置文件(存路径在python脚本提交目录下，即安装目录调试**步骤三**)** | **研发提供脚本**  **vi init.cfg**  根据脚本脚本中的参数定义、以及远程主机等信息修改init.cfg变量维护信息。 |

## 部署代码清单

|  |  |  |  |
| --- | --- | --- | --- |
| **程序编号** | **程序名称** | **处理周期** | **程序说明** |
| 1 | mda\_plan\_create\_table\_yyyymm.py | 月 | 用户历史行为信息、套餐维表、B域用户数据表、用户信息表创建 |
| 2 | mda\_plan\_list\_export.py | 月 | 向数据挖掘引擎提供数据文件和任务文件 |
| 3 | mda\_relevance\_plan\_file\_check.py | 日 | 下载并加载下个月用户预测信息 |
| 4 | mda\_plan\_userfitnew\_yyyymm.py | 月 | 适配结果用户清单 |
| 5 | mda\_plan\_userscor\_yyyymm.py | 月 | 用户套餐打分 |
| 6 | mda\_plan\_hidlist\_yyyymm.py | 月 | 得到用户潜在清单 |
| 7 | mda\_plan\_knn\_yyyymm.py | 月 | 4G套餐对非4G套餐的Knn算法计算相似度矩阵 |
| 8 | mda\_plan\_simily\_yyyymm.py | 月 | 解析Knn算法输出数据得到4G对非4G的相似度矩阵 |
| 9 | mda\_plan\_data\_analysis\_yyyymm.py | 月 | 导出样本数据和校验数据 |
| 10 | mda\_plan\_ruleprocess\_itemadd\_yyyymm.py | 月 | 经数据挖掘，解析之后得到用户套餐清单 |
| 11 | mda\_plan\_listhz\_yyyymm.py | 月 | 数据汇总 |
| 12 | mda\_plan\_recomd\_handler\_yyyymm.py | 月 | 套餐推荐handler（这个是前台点击推荐后，会调用到后台的py监控服务，然后监控服务会调用到此脚本，后台测试时次脚本可略过，直接执行脚本13） |
| 13 | mda\_plan\_recommend\_action\_yyyymm.py | 月 | 套餐推荐（前台点击推荐之后掉用的具体的推荐脚本的逻辑） |
| 14 | mda\_plan\_index\_yyyymm.py | 月 | 效果评估 |

## 实施人员注意事项

1. 由于settings.py中带有明文的用户名、密码，部署调试完成后请编译成pyc文件并删除.py文件，以防止用户名及密码泄露。
2. 在init.cfg 中配有ftp、及scp用户名、服务器ip、数据保存路径，需要根据本地的文件存储规划等相关规定进行修改，具体请查看脚本中的参数定义，这里的路径与数据挖掘平台配置的路径一致，配置的时候请注意。
3. 数据挖掘机器存放路径需要自行创建，详细路径格式参考init.cfg文件配置。

# 部署验证

1.脚本验证，保障脚本能正常运行，空测

python mda\_plan\_create\_table\_yyyymm.py 201510

2.ftp验证

首先验证

ftp ip（接口机）

输入用户名

输入密码

测试是否登录成功，如果不成功，告知本地开通FTP

3.挖掘验证

在数据目录下上传测试数据文件time\_201511.txt，任务文件20151130\_time\_201511\_machinelearning\_task.txt

然后执行

java -jar dataminer.jar 2 2 /home/ocdc/tas/shiting/tempData /home/ocdc/tas/guojy/planRes R 5 /home/ocdc/tas/shiting/data default

查看20151130\_time\_201511\_machinelearning\_task.txt是否 变成20151130\_time\_201511\_machinelearning\_task.txt.old

监控挖掘日志

查看挖掘结果目录下是否生成xx.fdm

# 部分脚本调试参考

1. scp 任务文件和数据文件上传验证

后台执行脚本nohup python mda\_plan\_list\_export.py 201511 >logs/ mda\_plan\_list\_export\_20151112.log &

执行完之后，查看 挖掘机器上/tempData（mkdir）目录下，是否有格式为

20151130\_time\_201511\_machinelearning\_task.txt任务文件和time\_201511.txt数据 文件存在，并查看其内容是否正确

2、数据挖掘验证

切换到 Root 用户，在挖掘文件安装所在目录下即：maching\_learning\_main.sh所在目录，执行数据挖掘调度命令。

Eg：

java -jar dataminer.jar 2 2 /home/ocdc/tas/shiting/tempData /home/ocdc/tas/guojy/planRes R 5 /home/ocdc/tas/shiting/data default

查看 挖掘机器上/data 目录下。是否输出 格式为data/time\_201511/ time\_201511\_R\_default\_txt\_20151130.fdm 结果文件，并查看其内容是否正确。

3、ftp下载验证

后台执行脚本 python mda\_relevance\_plan\_file\_check.py 201510 >logs/ mda\_relevance\_plan\_file\_check\_20151012.log &

执行完之后，查看本地文件是否存在。

# 数据核查办法

进入hive库，通过select 语句查询结果输出表，可以从一下几个方面查看数据：

1，数据源核查  
   核对维度是否超出所提供的维表数据范围  
   字段个数、字段类型、分隔符是否与接口规范一致  
   是否存在空值记录  
   是否有重复数据  
2，脚本结果数据核查

|  |  |  |
| --- | --- | --- |
| **程序结果数据核对** |  |  |
| 程序名称 | mda\_plan\_userfit\_yyyymm.py | |
| 核对指标 | 每个套餐的用户数 | |
| 源表： | 目标表：mda\_plan\_hidlist\_yyyymm | |
| mda\_plan\_userscor\_yyyymm |
| mda\_plan\_userdata\_yyyymm |
| 源表核查sql | 目标表核查sql | |
| select count(distinct tb.phone\_no) | select  count(distinct phone\_no) ,plan\_id from mda\_plan\_hidlist\_yyyymm where month\_id ='201510' group by plan\_id ; | |
| ,tb.plan\_id |
| from |
| ( |
| select b.phone\_no |
| ,b.plan\_id |
| from mda\_plan\_userscor\_yyyymm b |
| where month\_id ='201510' |
| ) tb |
| left outer join |
| ( |
| select a.plan\_id |
| ,a.phone\_no |
| from mda\_plan\_userdata\_yyyymm a |
| where month\_id ='201509' |
| ) ta |
| on (ta.phone\_no=tb.phone\_no) |
| where ta.plan\_id <> tb.plan\_id |
| 核查结果：  87 10001 89 10002 38 10004 55 10005 61 10006 93 10009 2 10010 31 10011 | 核查结果：  87 10001 89 10002 38 10004 55 10005 61 10006 93 10009 2 10010 31 10011 | |
|
|
|
|
|
|
|
| 核对结果是否一致 | 是 | |

# 常见问题处理

1.问题描述：scp到数据挖掘服务器上的数据文件乱码

解决：hive.exec.compress.output =false