生产平台多租户模型部署

使用须知

用户画像项目合作

中国联通

大数据公司-生产多租户

## 声明

1、模型生成文件个数、大小和存储周期要求：租户日模型产生文件个数原则上不超过1500个，每个文件应至少是兆级以上的文件，每天生成文件总量不超过200G（不含备份存储）；月模型产生的文件个数不超过5000个，每个文件应至少是兆级以上的文件，每月文件存储不超过400G（不含备份存储）。中间结果文件默认保留5天，结果文件默认保留7天，如需特殊延长保留周期，请在该部分特别说明。如保存周期大于30天或者文件存储超过限制，需要进行商务谈判。

**特别说明样例：**XXX模型中间结果需要保留 XX天，最终结果需要保留 XX 天。

2、根据在能力开放平台测试的情况，请在以下表格中写明模型预估产生的文件个数、预估的文件大小和数据保存周期，包括模型的中间结果。

|  |  |  |
| --- | --- | --- |
| 预估项 | 个数或者大小 | 备注 |
| 模型最终结果大小 | XX G/天 |  |
| 模型最终结果文件个数 | XX个/天 |  |
| 模型中间结果大小 | XX G/天 |  |
| 模型中间结果文件个数 | XX个/天 |  |
| 缓存结果大小 | XX G/天 |  |
| 缓存结果文件个数 | XX个/天 |  |

3、该部分写明模型迭代周期：模型正式上线一个月后，大模型和小模型相加，每月迭代不超过3次；模型小迭代的范围包括小范围更换配置文件和不修改模型逻辑的jar包替换；大迭代范围包括整个模型优化、输出字段调整和配置文件的大量变更。

4、因模型训练时能力开放平台为部分样例数据，实际生产平台数据量会比较大，请客户在生产平台进行测试前，做好模型的优化工作，如合并小文件（合并处理后1G~2G）和map join的性能优化等。

5、为增强模型的可移植性和可配置性，涉及到hive表的库名统一设置为参数。

6、用户离网的上网日志数据，因为需要进行数据初筛工作，所以数据提供周期为前一天的数据。

7、关于模型测试所需时间的说明：

1）对。于未按照《生产平台多租户模型部署说明》要求出现的问题，反馈时间一般为12工时 ；

2）对于模型部署出现的问题，首次。反馈时间一般为T+2个工作日（T日为部署说明中的问题已经解决的时间）

8、关于上线部署的说明：上线部署中，若涉及到互联网日志，在2个工作日内可部署完成，不涉及互联网上网日志在1个工作日内可部署完成。

9、.考虑数据交付及时性问题，日模型结果数据如果涉及精准营销平台，运行时间不能超过12小时自动终止，月模型加工不能超过48小时，超过则模型加工自动终止。

10、集群环境的版本

Java：1.8.0\_151-b12

Spark：2.2.1

Python： 2.7.5

Scala：2.11.8

Hive：1.1.0-cdh5.5.4

Hadoop 2.6.0-cdh5.13.1

## 二、部署规范

1、互联网日志范例

186xxxx9253|650285|11|868520034506905|900|2018-05-28 15:19:46.1188560|2018-05-28 15:24:46.4071180|29177

|544|164|708|6|10.103.51.40|223.167.105.44|0||3gnet.mnc001.mcc460.gprs|460015374621091|10.100.206.150|10.100.33.1||55318|443|1|5|

186xxxx2387|33937|31|357710060205054|912|2018-05-28 15:19:51.9664070|2018-05-28 15:19:53.4129400|276|124

|156|280|6|10.136.27.167|202.97.224.68|0||3gnet.mnc001.mcc460.gprs|460014640632148|116.79.237.150|220.206.138.129||21607|53|1|2|lm4.uuserv70.com

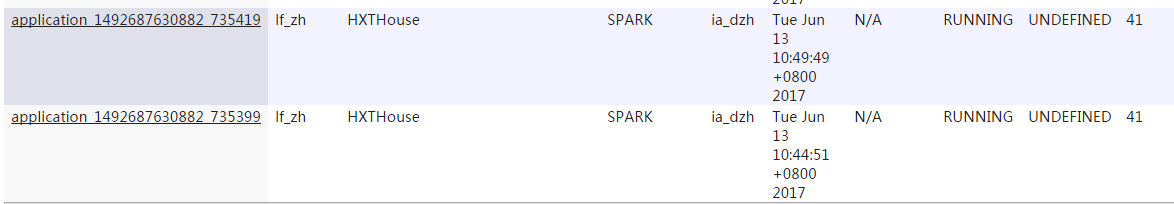
132xxxx8568|650519|31|864086035480577|915|2018-05-28 15:19:51.9754320|2018-05-28 15:20:01.9912080|10015

|132|0|132|6|10.24.20.209|202.97.224.69|0||3gwap.mnc001.mcc460.gprs|460011645204382|10.101.176.246|10.100.33.1||62464|53|0||accscdn.m.taobao.com

2、程序job\_name规范

请在编写程序是将的程序job\_name命名为租户名@模型名称格式，以便后续进行程序监控是能够对各个租户的模型加以区分

例如：如慧信通的房产模型job\_name命名为：[hxt@House](mailto:hxt@House)



3、输出数据分隔符

租户模型输出数据请全部设置为'|'分隔符

样例数据如下：



4、租户模型参数设置

模型程序中涉及到输入目录、配置文件、输出目录、hive数据库名、执行账期、省份和程序执行队列，请设置为可执行jar包外的参数，方便部署人员配置；

5、模型资源配置

预计3-4节点（每节点目前为24C+80G内存）

1. **常见问题汇总**

**1、模型输入数据**

1. 数据规范中字段与数据可开放数据规范不符合 例如：流量上网日志原始表生产平台中共有26个字段，租户输入字段超出26个字段范围；
2. hive表中数据以hdfs方式进行数据输入，读取分区字段，实际文件存储中分区字段是不存在的；
3. hive表直接作为模型输入以select方式读取时，字段名称与生产平台规范不符合；
4. 流量上网日志需要进行筛选的，在部署文档提交的同时请提交相关筛选规则；
5. 上网日志筛选的host列表文件，需在文档中注明，请避免出现筛选规则找不到的情况；
6. 部署文档输入字段与模型实际输入字段不符；
7. 租户跑取全国数据传入参数为099，实际生产上并无此省分字段，需指代明确或者参数以0\*代指全国；
8. 输入数据不是可开放数据规范里的原始数据的，需在程序里写明加工脚本，涉及的数据存放路径以参数传入即可，部署人员不做二次加工或定制化数据需求；
9. 租户自己的输入数据需与生产数据保持一致，如均为加密格式，否则无法与生产数据关联匹配。

**2、模型输出数据**

1. 模型输出字段与部署文档不符
2. 模型输出数据请明确写明是分省数据还是全国数据，方便部署人员进行后续数据推送
3. **模型部署时问题**
4. 文档配置文件需修改处，请在部署文档中写明，做出重点标注
5. 执行脚本样例请写明在部署文档中，不要单独写在文档外部
6. 执行程序的命令及样例在部署文档中填写规范，尤其涉及到主类和参数。请不要单独在文档外填写；
7. 租户模型依赖python的依赖包，生产部署测试只能client模式测试，无法用cluster集群模式跑取，所需的依赖环境需自行封装成包进行调用。
8. 租户跑取的结果数据需直接写入hdfs或者hive表里，所需的路径或者表名建议传参数或者统一按规范取名，例如库名为，lf\_xhsy\_pro，中间为租户名称缩写。
9. 租户的建表语句等需按部署文档规范命名并提交，不要按照测试时命名。

**四、模型迭代更新**

模型迭代时必须同时更新部署文档，请在原部署文档基础上填写迭代内容