**中国移动统一AI平台**

**能力入驻部署手册**

2020-7



**修订记录**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 版本号 | 修订日期 | 修改人 | 修改内容 | 修改原因 |
| V0.1 | 2020-07-23 | 付上海 | 无 | 初稿 |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

**目录**

Contents

[一、 入驻概要流程 4](#_Toc46405077)

[二、 角色及联系人 5](#_Toc46405078)

[三、 详细步骤 5](#_Toc46405079)

[1. 入驻申请 5](#_Toc46405080)

[2. 接入方案评审 6](#_Toc46405081)

[3. 账号申请 6](#_Toc46405082)

[4. 账号验证 6](#_Toc46405083)

[4.1 VPN账号登陆 7](#_Toc46405084)

[4.2 4A账号登陆 8](#_Toc46405085)

[4.3 统一AI平台登陆 9](#_Toc46405086)

[5. AI能力部署/发布 11](#_Toc46405087)

[5.1 镜像上传 11](#_Toc46405088)

[5.2 能力部署 12](#_Toc46405089)

[5.3 能力发布 15](#_Toc46405090)

[6. 推理服务发布 16](#_Toc46405091)

[6.1 创建分组 17](#_Toc46405092)

[6.2 创建API接口 17](#_Toc46405093)

[6.3 接口授权 18](#_Toc46405094)

[6.4 获取应用密钥 19](#_Toc46405095)

[7. 业务方调用 20](#_Toc46405096)

[7.1 内网访问统一AI平台服务 20](#_Toc46405097)

[7.2 外部业务访问统一AI平台服务 20](#_Toc46405098)

[四、 参考资料 21](#_Toc46405099)

# 入驻流程



# IT公司及技术支持联系人

|  |  |  |  |
| --- | --- | --- | --- |
| 姓名 | 单位/部门 | 职位/职责 | 联系方式 |
| 陶涛 | IT公司 | IT公司副总（分管领导） | - |
| 尚晶 | IT公司 | IT公司处长 | - |
| 孙昊 | IT公司 | IT公司大数据技术部接口人 | 139-1111-3353  [sunhao@chinamobile.com](mailto:sunhao@chinamobile.com) |
| 王瑞 | 科大讯飞 | 技术支持团队接口人 | 186-1233-5614  [ruiwang24@iflytek.com](mailto:ruiwang24@iflytek.com) |
| 刘云振 | 科大讯飞 | 技术支持团队技术支撑 | 181-5887-2718  [yzliu10@iflytek.com](mailto:yzliu10@iflytek.com) |

# 详细步骤

## **入驻申请**

由接入方项目经理填写如下3个申请资料，邮件发送至IT公司大数据平台部接口人（孙昊 [sunhao@chinamobile.com](mailto:sunhao@chinamobile.com)）

1. 《统一AI平台租户入驻申请表》-**该申请表需TSG部门领导签字，并盖部门公章**

说明：对于平台资源（GPU卡、CPU、内存、存储空间等）申请，需结合业务量及模型单元处理能力给出资源申请依据依据。如：AI模型单GPU卡可处理请求1000次/秒，并发业务量6000次/秒，则申请6000/1000=6块GPU卡

模板示例如下：



注意：请勿遗漏申请项（如：存储卷）导致重新申请影响入驻进度。

1. 《AI项目介绍&建设方案》



1. 《AI场景能力分解表》



## **接入方案评审**

由IT公司大数据平台部组织接入方及技术支持团队讨论接入方案，评审通过启动接入工作，如评审不通过则需要接入方（咪咕公司）修订。

## **账号申请**

账号申请包括集团VPN，4A及统一AI平台账号，需提供如下3个资料邮件发送至技术支持团队接口人（王瑞[ruiwang24@iflytek.com](mailto:ruiwang24@iflytek.com)）。

1. 《网络与信息安全责任承诺书-企业和个人》

其中合作方企业承诺书需在授权代表处由**咪咕VP签字并盖部门工章**；合作方员工承诺书需账号申请人员及**咪咕VP签字，且盖部门公章**, 以扫描件的方式发送至

1. 《大数据平台合作伙伴人员变动反馈-模板》



1. 《安全意识培训考试试题》

微信扫描如下二维码，第五套为必答，其他四套再选择一套答题即可，合格标准：为90分（含）以上，若不合格则需要重考，考试不限次数；考试合格后将考试分数截图。

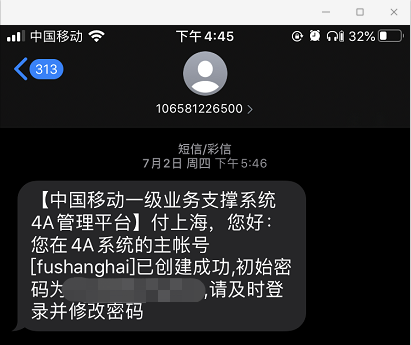
|  |
| --- |
| 第一套：第二套：第三套：  第四套： 第五套（必答）：C:\Users\cmsz\AppData\Local\Temp\WeChat Files\f559a762ca081c2a29dd8c3a7efe17b.jpg |

## **账号验证**

账号登陆顺序为：VPN登 -> 4A系统登陆 -> 登陆图形堡垒机 -> 堡垒机登陆统一AI平台。

### **VPN账号登陆**

待技术支持团队创建后，VPN及4A账号以短信方式通知，如图所示：



浏览器登陆集团VPN下载网站下载VPN客户端： <https://117.136.128.74/welcome.php>



安装完成后开客户端登陆使用短信中的4A账号登陆（VPN账号与4A账号一致），VPN服务器地址：117.136.128.74

登陆成功出现右下角提示，表示VPN登陆成功。



### **4A账号登陆**

VPN连接后浏览器访问4A系统 ，输入4A账号(账号及密码通过短信获取)并通过短信获取验证码进行登陆：

4A网址：<http://10.255.246.33:7080/uac/web3/jsp/login/login.jsp>

注意：必须使用IE浏览器登陆



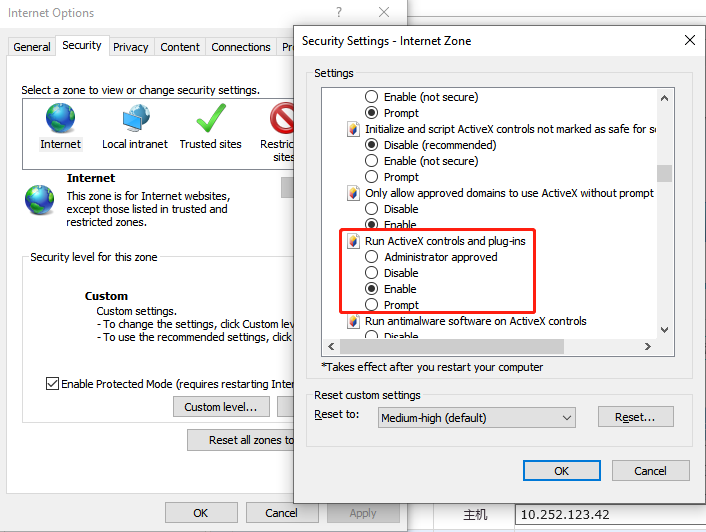
登陆后系统如下



### **统一AI平台登陆**

使用IE浏览器登陆4A系统前，确保IE浏览器运行ActiveX运行已打开。

IE设置 -> Internet选项 -> 安全选项卡 -> 自定义级别



登陆4A系统后，进入“设备登陆”，随机选择一台图像堡垒机，点击“登陆图标”

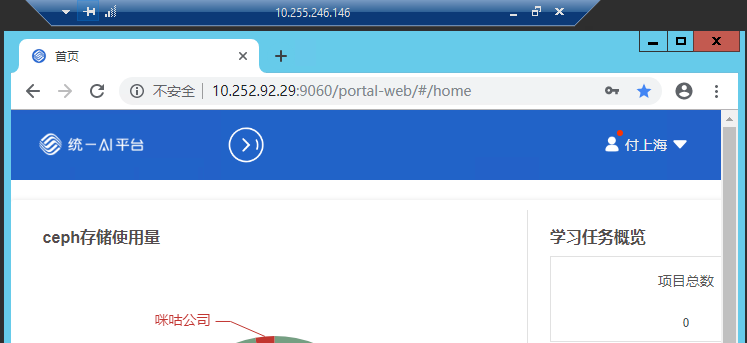


登陆图像堡垒机（远程桌面）后，Chrome浏览器中访问<http://10.252.92.29:9060/portal-web/#/home>

注：**账号与4A账号一致，默认密码为123abc，首次登陆后必须更改**。



登陆成功后出现如下所示界面：



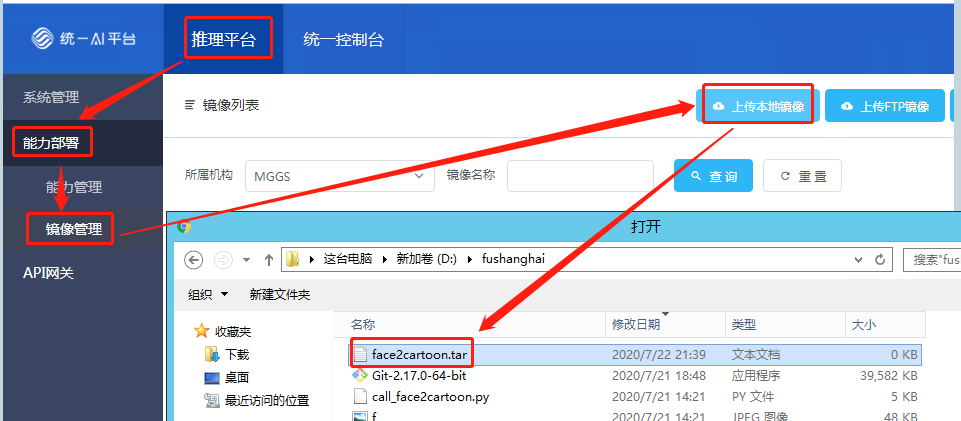
## **AI能力部署/发布**

### **镜像上传**

方法一：在图形堡垒机使用平台界面上传镜像文件（适用与小于4G镜像）

步骤1：用户将镜像文件上传至图形堡垒机（远程桌面拷贝）

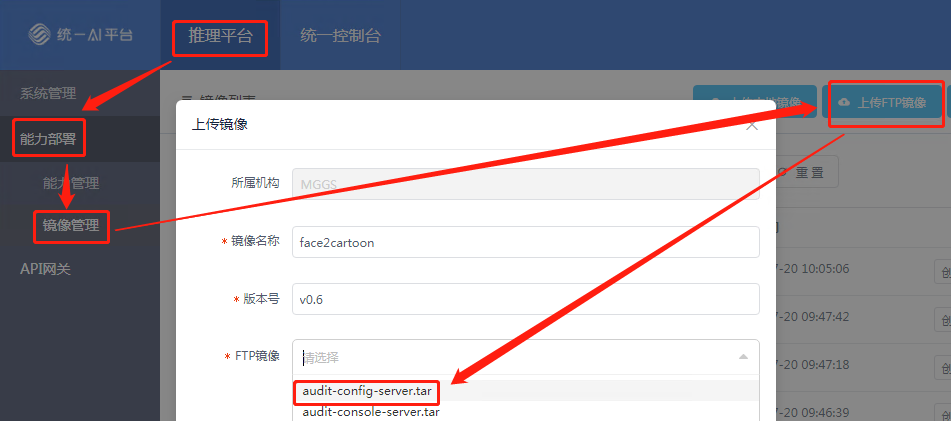
步骤2：用户进入统一AI平台，选择【推理平台】->【能力部署】->【镜像管理】->【上传本地镜像】，将本地镜像上传至机构指定Harbor仓库中。



方法二：通过FTP上传镜像文件（适用与小于4G镜像）

步骤1：将镜像文件打成tar包，通过中国移动和彩云网盘方式发给平台技术支持团队上传。

步骤2：进入推理平台，选择【推理平台】->【能力部署】->【镜像管理】->【上传FTP镜像】，将FTP服务器中镜像tar包上传至机构指定Harbor仓库中。



### **能力部署**

进入推理平台，选择【能力部署】->【能力管理】-> 【新增】



选择指定镜像，输入“项目名称“和”项目描述“后，完成部署项目创建。



编辑部署信息，进入推理平台，选择【能力部署】->【镜像管理】，点击能力列表后的部署信息，完成对项目的部署信息编辑





* 基本信息：根据业务需求设置基础资源，并添加存活探针、应用监控等服务代码。
* CPU单位：毫核 GPU单位：卡 内存单位：M
* 配置文件：如不输入挂载目录，系统默认为/app/cfg，可选择此配置文件进行编辑。
* 环境变量：设置K/V形式的环境变量参数。
* 目录挂载：设置挂载所需目录，支持host和pv（k8s集群提供）。
* 添加服务：按需发布成服务，支持Cluster IP和NodePort2种方式。
* 服务代理：依据业务需求设置服务代理来实现外部访问。
* 自动伸缩：在机构资源限额下，根据容器对内存、CPU request使用率设置最大/最小实例数。

### **能力发布**

进入推理平台，选择【能力部署】->【能力管理】，点击能力列表后“发布“，完成对AI能力项目的发布。



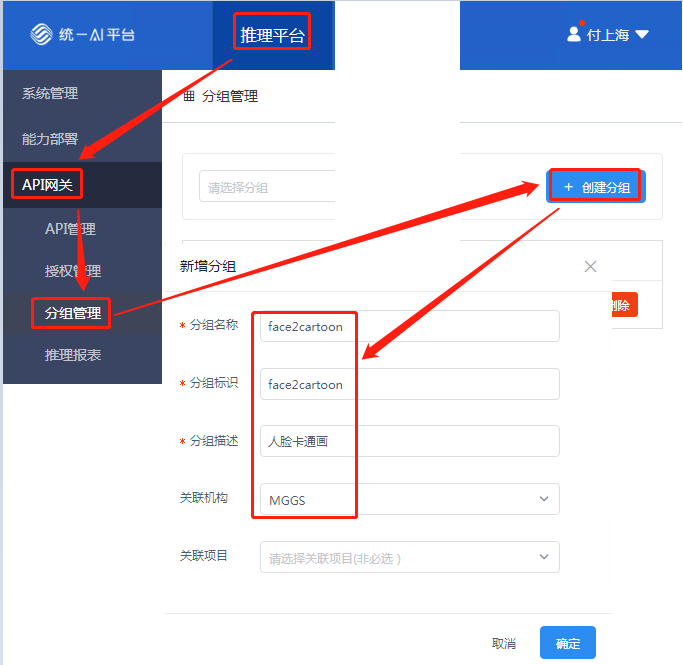
点击发布后可查看发布状态



如发布失败，可在发布页面查看【发布历史】，点击【查看失败信息】

## **推理服务发布**

### **创建分组**

进入推理平台，选择【API网关】->【API网关】-> 【分组管理】-> 【创建分组】，下一步骤将在该分组下创建API接口

### **创建API接口**

进入推理平台，选择【API网关】->【API管理】-> 【创建API】，选择新创建的分组，填写API Path地址及名称，版本等信息。



### **接口授权**

接口授权按组为单位，推理服务为某分组下所有API接口授权添加API。

进入推理平台，选择【API网关】->【授权管理】-> 【新增应用】，填写应用标识和应用名称。



进入授权应用列表，点击【编辑授权】，在弹出页面中将新创建的分组加入到该应用中。



### **获取应用密钥**

进入推理平台，选择【API网关】->【授权管理】在应用列表中选择【编辑密码】



获取如下密钥信息，该应用标识和密钥需妥善保存，并提供给业务调用方，当业务方访问本AI能力时以此密钥做校验，无正确密钥将拒绝访问。



## **AI能力测试验证**

## **业务方调用**

### **内网访问统一AI平台服务**

用户通过图形堡垒机访问统一AI平台能力。

访问URL：统一AI平台内网IP地址：统一AI平台服务端口/API path

URL说明：

1) 统一AI平台内网IP地址（固定地址）： 10.252.92.29

2）统一AI平台服务端口（固定端口）：9050 API调用

3）API path：在推理服务发布中创建的API path

示例：<http://10.252.92.29:9050/facecartoon/v1/request>

### **外部业务访问统一AI平台服务**

用户通过资源池内部、承载网、公网访问统一AI平台访问AI能力。

访问URL：统一AI平台资源池私网IP/承载网/公网IP地址：统一AI平台服务端口/API path

URL说明：

1) 统一AI平台资源池私网IP地址（固定地址）： 10.252.122.96（业务服务器在核心区内）

2）统一AI平台承载网IP地址（固定地址）： 10.255.77.9

3）统一AI平台公网访问地址（固定地址）：117.132.181.235

4）统一AI平台服务端口（固定端口）：9050 API调用

5）API path：在推理服务发布中创建的API path

示例：<http://10.255.77.9:9050/facecartoon/v1/request>

# 参考资料

详细平台操作录屏：<http://bigdata.10086.cn/artint/px/>