

腾讯云云智天枢集群部署规划指引

使用手册

[2020.03]

[V2.0]

**【版权声明】**

**©2017-2018 腾讯云版权所有**

本文档著作权归腾讯云单独所有，未经腾讯云事先书面许可，任何主体不得以任何形式复制、修改、抄袭、传播全部或部分本文档内容。

**【商标声明】**



及其它腾讯云服务相关的商标均为腾讯云计算（北京）有限责任公司及其关联公司所有。

本文档涉及的第三方主体的商标，依法由权利人所有。

**【服务声明】**

本文档意在向客户介绍腾讯云全部或部分产品、服务的当时的整体概况，部分产品、服务的内容可能有所调整。您所购买的腾讯云产品、服务的种类、服务标准等应由您与腾讯云之间的商业合同约定，除非双方另有约定，否则，腾讯云对本文档内容不做任何明示或模式的承诺或保证。

**目录**

[1 前言 6](#_Toc22289457)

[2 云智天枢简介 6](#_Toc22289458)

[2.1 功能架构 6](#_Toc22289459)

[2.2 组件介绍 8](#_Toc22289460)

[2.3 节点角色 9](#_Toc22289461)

[3 云智天枢硬件规划 10](#_Toc22289462)

[3.1 网络选型 10](#_Toc22289463)

[3.2 服务器选型 11](#_Toc22289464)

[3.2.1 服务器硬件要求 11](#_Toc22289465)

[3.2.2 内存部署规划 13](#_Toc22289466)

[3.2.3 硬盘部署规划 13](#_Toc22289467)

[3.3 服务器数量评估 16](#_Toc22289468)

[4 云智天枢软件规划 18](#_Toc22289469)

[4.1 操作系统要求 18](#_Toc22289470)

[4.2 组件拓扑规划 18](#_Toc22289471)

[4.3 资源消耗型组件和服务的合理分布 18](#_Toc22289472)

[5 云智天枢典型应用场景部署案例 18](#_Toc22289473)

**文档修订记录**

|  |  |  |  |
| --- | --- | --- | --- |
| **版本** | **修改日期** | **修改人员** | **说明** |
| V1.0 | 2019.10 | kachidokima | 集群规划初稿 |
| V1.1 | 2019.10.12 | aronhlchen | 调整格式和不明确的部分 |
| V2.0 | 2020.03.17 | kachidokima | 添加监控组件说明与资源评估、es存储资源评估、网关team绑定要求 |

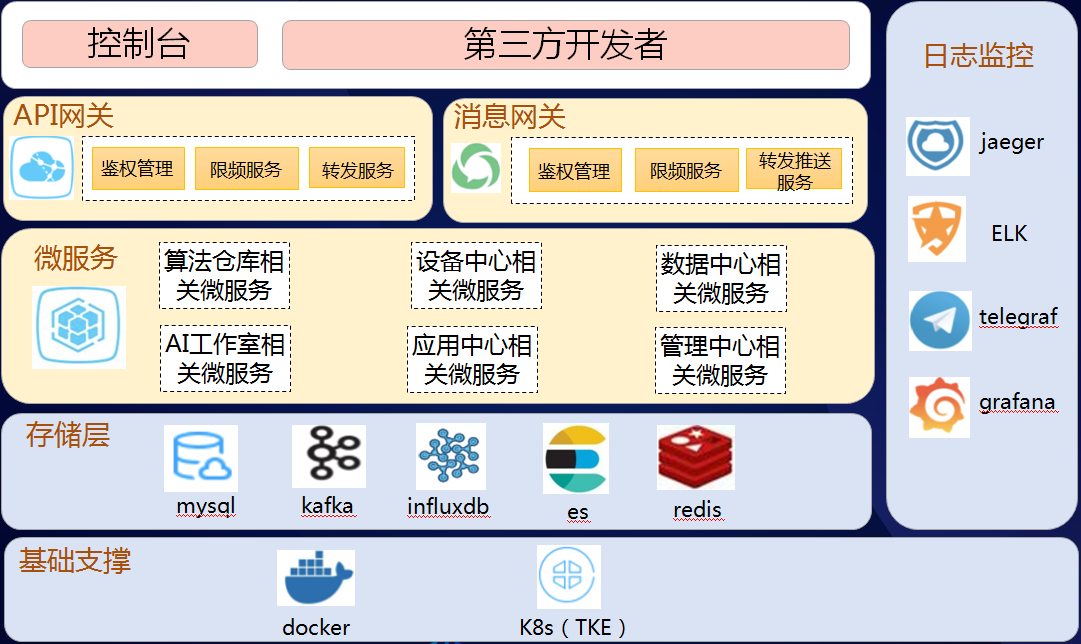
1. 前言

腾讯云智天枢人工智能服务平台（TI Matrix Platform，下文中也叫云智天枢平台）是基于腾讯云和腾讯优图强大技术能力打造的全栈式人工智能服务平台，致力于帮助加速企业的数字化转型及促进 AI 行业生态共建。平台支持快速接入各种数据、算法和智能设备，并提供可视化编排工具进行服务和资源的管理及调度，进一步通过 AI 服务组件集成和标准化接口开放，降低 AI 应用开发成本。

本文档从云智天枢平台建设的部署规划入手，通过网络、硬件选择、以及功能组件选择和应用场景说明，帮助客户在云智天枢平台建设前尽可能做好规划，方便部署过程中选择安装对应的服务组件和后续平台扩展。

1. 云智天枢简介

## 功能架构

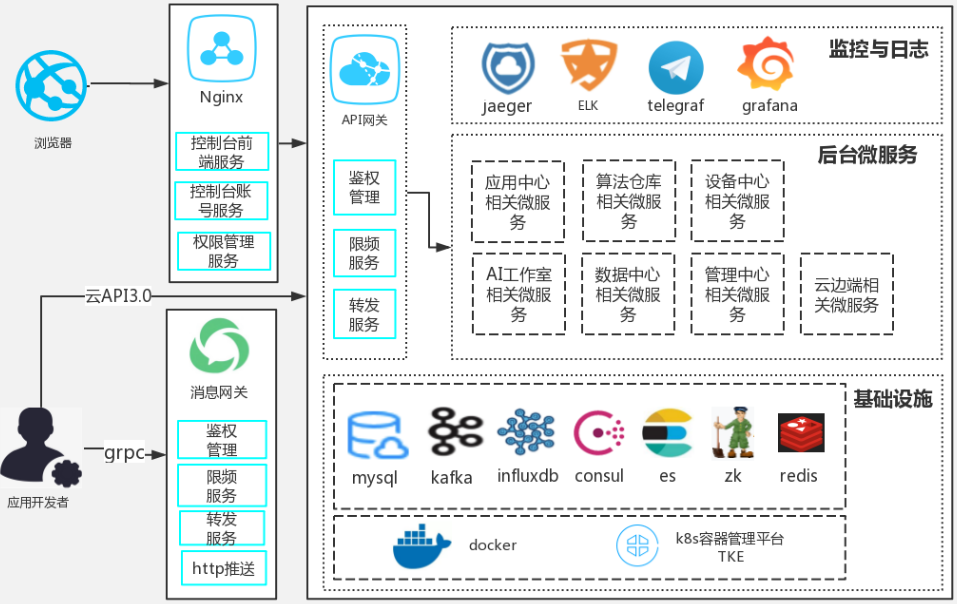


如图所示，云智天枢人工智能服务平台是主要分为以下几个部分。

* 基础设施：位于整个架构的最下层，负责整体平台基础性能力的支撑，如 Docker、 K8S等。
* 存储层：负责存储和管理平台服务的数据，用到了一些诸如 MySQL、Kafka、Influxdb、ES 等组件。
* 微服务：平台的主体。根据不同的功能，平台被划分为6个主要的模块，分别是算法仓库、设备中心、数据中心、AI工作室、应用中心以及管理中心。
* 算法仓库：主要提供自助打镜像的能力，可快速把可执行程序、模型文件等容器化为算法微服务等，目前接入算法种类50+，涵盖人脸，车辆，语音，文字，语义等。
* 设备中心：主要对接各个厂商的各个型号的设备，比如普通摄像机，抓拍机，AI相机等等，支持设备厂商自助设备接入的能力。
* 数据中心：主要负责数据接入、推送、转换、存储等，包括屏蔽各种结构化与非结构化存储介质的能力。
* AI工作室：主要实现了任务调度，流程与服务的编排能力，打通平台各个窗口的能力。
* 应用中心：主要提供创建应用、密钥、订阅管理，视图库等能力。
* 管理中心：主要提供账号系统、角色权限、镜像仓库、操作日志等能力。
* 网关层面：分为API网关以及消息网关两部分。
* API网关：采用的是 腾讯云API3.0 标准，主要做鉴权、限频、转发等功能；
* 消息网关：支持GPRC和http推送能力。
* 监控系统：用了 Telegraf 、InfluxDB和Grafana，日志系统采用的是 ELK。

## 组件介绍

腾讯云智天枢平台组件架构：



云智天枢用的组件包括必要组件和监控组件，其主要功能和部署规划指引说明如下：

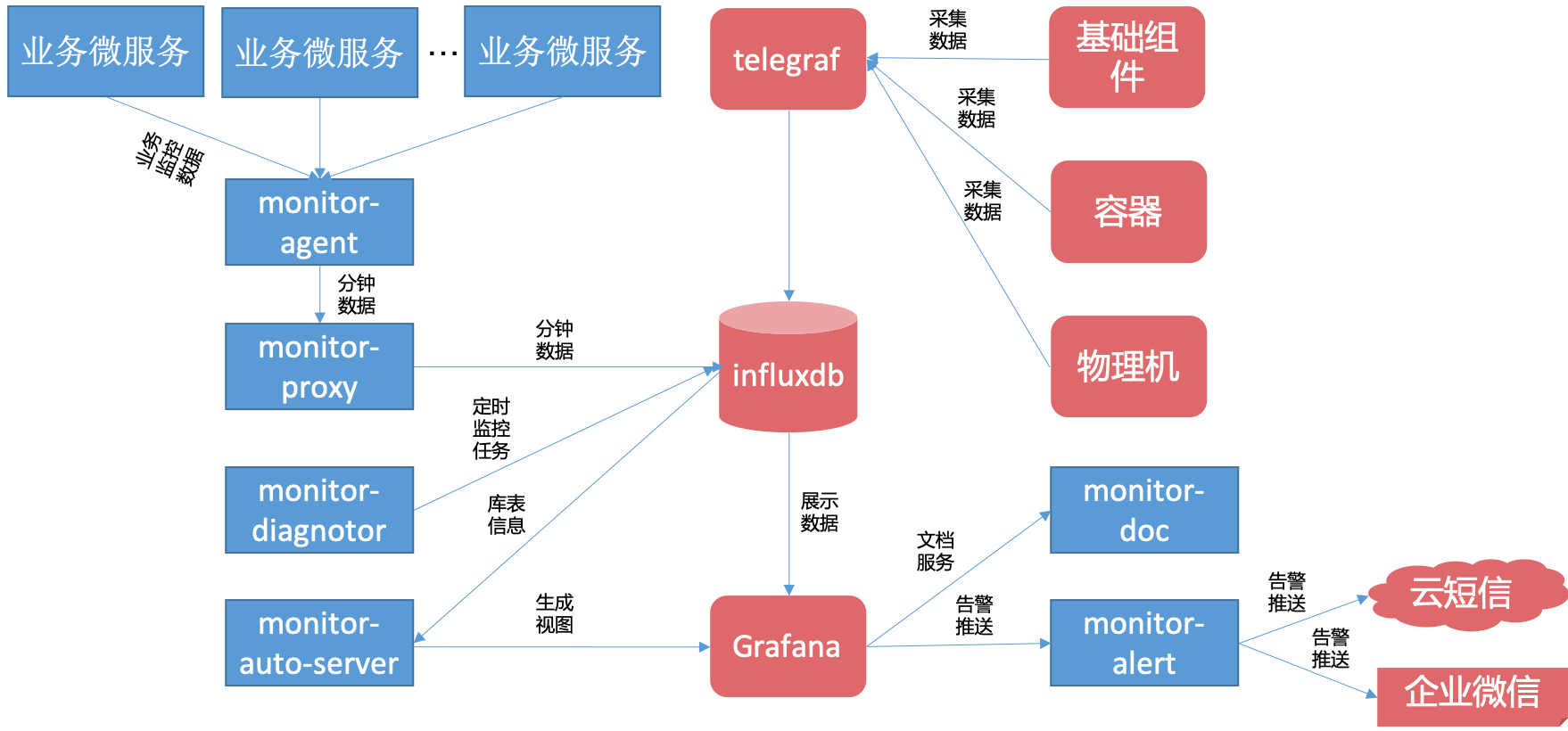
* 必要组件
* Zookeeper：用于服务注册、配置中心。集群模式至少部署3个节点（奇数个），1 leader，多follower；
* MySQL：结构化信息永久存储。集群模式为1主1从；
* Consul：组建集群，管理组件域名与健康检查。集群模式至少部署3个节点，无master；
* Redis：存储后台的数据信息。集群模式为1主1从；
* ES: 分布式收集并存储海量数据，集群模式至少部署3个节点，无master；
* Kafka：消息中间件，用于服务间通信与消息订阅。集群模式至少部署3个节点，无master；
* 监控组件：
* Influxdb：存储时序数据，主要用于监控数据存储。部署在其中一个节点即可，注意需要与Grafana部署在相同节点；
* Telegraf：采集监控数据。所有节点均需部署；
* Grafana：将采集的数据可视化，部署在其中一个节点即可，注意需要与Grafana部署在相同节点。

注意：在安装时部署角色可能不单是组件还会包含平台服务和初始化操作，具体角色和解释参考《部署文档》-3.2.1安装组件包

在实际部署时（例如公有云场景）有些组件资源不需要安装，直接复用客户组件或者云上业务组件，这种场景可以在安装时跳过一些组件安装，具体参考《部署文档》-3.2.1安装组件包

## 监控系统架构

云智天枢监控系统架构如下图：



监控系统通过开源组件telegraf 、influxdb 、grafana 搭建。其中：

* telegraf用于采集容器与物理机还有基础组件信息，每台机器均会部署；
* influxdb与grafana均为单机部署；
* monitor-agent会采集每个微服务的数据，采用容器化daemonset部署的方式；
* monitor-diagnotor、monitor-proxy、monitor-auto-server、monitor-doc、monitor-alert均采用deploy的方式容器化部署。

所有数据最终都会存储到influxdb，然后由grafana来生成监控视图。

## 节点角色

一般情况下，云智天枢集群中的服务器节点可划分为以下几类角色，见表1：

注意：

1. 以下仅以高可用的场景为例进行说明，其他部署方式可参考《腾讯云智天枢平台实施要求》-03-服务器配置要求；
2. 节点角色以逻辑原则划分，实际物理部署场景可能有混布情况。

*表1-组件类型信息表*

|  |  |  |
| --- | --- | --- |
| 角色 | 节点数量 | 部署内容 |
| 容器平台管理节点 | 3 | 运行容器平台的管理服务 |
| 平台算法运算节点 | 1+ | 运行业务算法服务容器（根据具体业务选择扩容） |
| 组件运算节点 | 3 | 组件：zookeeper、kafka、consul、es、redis、mysql(条件允许可独立部署两台) |
| 存储节点 | 1+ | 分布式文件存储（集群机器数量较少时可在其他节点上混布） |

1. 云智天枢硬件规划

本章节对云智天枢平台的硬件规划进行简要说明，同时读者可阅读《腾讯云智天枢平台实施要求》中的相关部分，互为参考。

## 网络选型

集群不同服务器之间会进行大量的数据交互，因此，对服务器网卡、对网络传输速度、对上游交换机等都有较高的要求。同时，为了提高数据传输速度，集群需要尽量部署在同一个机房内网，尽量保证数据不跨网传输。

推荐生产环境为万兆网络作为集群的基础网络要求，测试环境可以是千兆网络。基本要求如下：

1. 首先，服务器网卡至少保证千兆网卡配置，生产环境推荐万兆网卡做teaming并设置模式active-backup保证主备容灾；
2. 其次，对于规模较小的集群（不大于20台服务器）内网交换机网口至少千兆网口；对规模较大的集群（至少20台服务器），推荐使用万兆网口交换机；
3. 最后，集群所有节点需要部署在同一个内网并尽量在一个网段，减少跨网部署造成的跨网数据传输。

## 服务器选型

在对云智天枢集群的服务器进行选型前，一般需要进行详细的应用场景分析和性能测试，从而全面理解集群工作负荷，才能相对准确的评估得到服务器配置方案。然而，通常集群是一个形态变化的系统，因此一般建议的是，在开始部署生产环境时，尽量采用中高配置、高容灾性和具备负载均衡能力的硬件设备。当使用多种体系组件的时候，硬件资源的使用将会是多样的, 需要系统管理员专注于资源配置。

### 服务器硬件要求

如前文所述，云智天枢集群主要分为以下节点（实际情况下可能存在混布情况）：

* 容器平台管理节点
* 平台算法运算节点
* 组件运算节点

注意：组件包含mysql、zookeeper、kafka、redis、consul、es等。其中，mysql的存储需要独立评估。

* 存储节点

每一类节点对于CPU、内存、网卡、系统磁盘、数据磁盘等的要求都有差别，但一般为了便于采购机器的规格统一，尽量减少机器差异，具体信息见表2：

*表2 硬件需求详细信息表*

|  |  |  |  |
| --- | --- | --- | --- |
| **节点类型** | **节点数量** | **推荐配置** | **说明** |
| 容器平台管理节点 | =1 | * CPU： >= 16核（逻辑核）； * 内存： >= 32 GB； * 网卡：生产要求万兆以上，测试要求千兆以上； * 系统盘：推荐SSD盘，其次SAS盘；容量>=100 G；要求必须做raid，推荐raid1； * 数据盘：推荐SAS盘；容量>=300G；数据盘推荐使用raid1或raid5。 | 运行容器平台管理服务，在集群机器资源紧张（如数量较少）的情况下，可考虑混布平台业务。  一个节点，用于poc，接受不容灾，机器资源有限情况使用。推荐3个节点，且分布在不同的机柜。 |
| =3 | * CPU： >= 16核（逻辑核）； * 内存： >= 32 GB； * 网卡：生产要求万兆以上，测试要求千兆以上； * 系统盘：推荐SSD盘，其次SAS盘；容量>=100 G；要求必须做raid，推荐raid1； * 数据盘：推荐SAS盘；容量>=300G；数据盘推荐使用raid1或raid5。 |
| 平台算法运算节点 | =1 | * CPU： >= 56核 * 内存： > =128GB； * 网卡：生产要求万兆以上，测试要求千兆以上； * 系统盘：推荐SSD盘，其次SAS盘；容量>=100 G；要求必须做raid，推荐raid1； * 数据盘：推荐SAS盘；容量>=300G；数据盘推荐使用raid1或raid5； * 显卡：>=2卡P40，推荐使用P40显卡。 | 运行平台业务服务与平台算法服务，算法运算节点视需求而定，高可用场景下推荐多节点部署 |
| >=2 | * CPU： >= 28核 * 内存： > =56GB； * 网卡：生产要求万兆以上，测试要求千兆以上； * 系统盘：推荐SSD盘，其次SAS盘；容量>=100 G；要求必须做raid，推荐raid1； * 数据盘：推荐SAS盘；容量>=300G；数据盘推荐使用raid1或raid5； * 显卡：>=2卡P40，推荐使用P40显卡。 |
| 组件运算节点 | =1 | * CPU： >= 16核； * 内存： >= 64 GB； * 网卡：生产要求万兆以上，测试要求千兆以上； * 系统盘：推荐SSD盘，其次SAS盘；容量>=100 G；要求必须做raid，推荐raid1； * 数据盘：推荐SAS盘；容量300+mysql存储（独立评估）+ es存储（独立评估）；数据盘推荐使用raid1或raid5。 | 运行基础组件，如果集群机器数量较少也会运行平台业务服务 |
| =3 | * CPU： >= 16核； * 内存： >= 64 GB； * 网卡：生产要求万兆以上，测试要求千兆以上； * 系统盘：推荐SSD盘，其次SAS盘；容量>=100 G；要求必须做raid，推荐raid1； * 数据盘：推荐SAS盘；容量300+mysql存储（独立评估）+ es存储（独立评估）；数据盘推荐使用raid1或raid5。 |
| =5 | 3台用户部署：zookeeper，kafka，redis，consul，es   * CPU： >= 16核； * 内存： >= 64 GB； * 网卡：生产要求万兆以上，测试要求千兆以上； * 系统盘：推荐SSD盘，其次SAS盘；容量>=100 G；要求必须做raid，推荐raid1； * 数据盘：推荐SAS盘；容量>=300G；数据盘推荐使用raid1或raid5。   2台用于部署mysql   * CPU： >= 16核； * 内存： >= 64 GB； * 网卡：生产要求万兆以上，测试要求千兆以上； * 系统盘：推荐SSD盘，其次SAS盘；容量>=100 G；要求必须做raid，推荐raid1； * 数据盘：推荐SAS盘；容量mysql存储（独立评估）+ es存储（独立评估）；数据盘推荐使用raid1或raid5。 |
| 存储节点  (文件存储) | >=1，在集群数量较多资源充分情况下 | * CPU： >= 16核； * 内存： >= 32 GB； * 网卡：生产要求万兆以上，测试要求千兆以上； * 系统盘：推荐SSD盘，其次SAS盘；容量>=100 G；要求必须做raid，推荐raid1； * 数据盘：推荐SAS盘；容量>= 100T（如果数据量很大，建议采用更高的配置，需要额外评估）；数据盘推荐使用raid1或raid5。 | 可选配置，只用作分布式文件存储 |

### 内存部署规划

实际场景内存一般不会成为瓶颈。可以参考[4.3小节](#_资源消耗型组件和服务的合理分布)，并根据表格2，在不同的场景选择不同的机器，至少采用最低配置的内存即可。

### 硬盘部署规划

表格2详细阐述了不同节点所需要的系统盘和数据盘类型，为了提高数据安全性、容灾可恢复特性，建议尽量采购的服务器能够支持硬件raid卡功能。

对于系统盘和数据盘而言：

* 系统盘：无论是哪一类节点，都推荐使用SSD或者SAS盘。系统盘为了考虑容灾和备份，要求至少需要做raid1。系统盘的最终容量，要求不低于100GB。
* 数据盘：不同节点的要求不同，具体说明如下：

1. 容器平台管理节点、平台算法运算节点：

数据盘尽量选用SAS盘，且考虑到一些系统元数据和业务数据的备份容灾，数据盘做raid1或者raid5。数据盘的最终容量，要求不低于300GB；

1. 组件存储节点：
2. 通过用组件存储评估（zookeeper、kafka、redis、consul）

数据盘的最终容量，要求至少不低于300GB。如果组件上包含mysql需要独立评估存储，见下一小节。

1. es存储评估：

集群日志搜集采用elk的方式，由于会通过es搜集整个集群内的日志形成索引，所以会导致部署es的机器会产生额外的存储开销，可以通过以下公式预估：

S = P \* D \* 1

其中：

* S：es索引数据存储大小，单位为G；
* P：结构化任务路数；
* D：日志保留天数，一般建议7天；
* 1：平均每路每天产生索引大小为1G；

1. mysql存储评估：

mysql需要存储大量的结构化数据，尽量选用SAS盘，且考虑到一些系统元数据和业务数据的备份容灾，数据盘做raid1或者raid5。

针对数据存储量更大的场景，配置的要求更高。以下公式可用于评估推算结构化数据的存储大小：

S= ( ( (P1 + P2 + P3) \* M \* 30 \* 10000 ) / 10000000 ) \* 5

其中：

* S：结构化数据存储大小，单位为G；
* P1：人脸路数；
* P2：机动车路数；
* P3：非机动车路数；
* M：月数；
* 30：表示一个月的同天数为30天；
* 10000：表示一路视频一天最大抓拍数据为10000条；
* 10000000和5：表示10000000（1千万）条数据占用存储大小为5 GB。

1. 存储节点：

在需要存储大量的文件数据和视频数据的业务场景下，需要采用分布式存储。推荐选用SAS盘，且考虑到业务数据的备份容灾，数据盘做raid1或者raid5。数据盘的最终容量，要求至少不低于100T。

以下公式可用于评估推算：

S= ( (P1 + P2 + P3) \* M \* 30 \* 10000 \* (50 + 800) ) / ( 1024 \* 1024 \* 1024 )

其中：

* S：数据存储大小，单位为T；
* P1：人脸路数；
* P2：机动车路数；
* P3：非机动车路数；
* M：月数；
* 30：表示一个月的同天数为30天；
* 10000：表示一路视频一天最大抓拍数据为10000条；
* 50：表示一张小图占用50 K；
* 800：表示一张大图占用800 K。需要与用户确认大图的大小，一般由根据视频的分辨率决定；
* 1024：存储单位换算定值，从K转化为单位T，需要除以（1024\*1024\*1024）。

同时考虑应用场景，若需要性能要求较高的场景，建议可以将数据盘替换为SSD或者SAS盘，总体上磁盘IO性能上是SSD>SAS>SATA。

对于磁盘挂载的目录，要求如下：

1. 由于docker目录（/var/lib/docker）和业务数据(/data/timatrix)会占用大量的存储，需要确保这两个目录在数据盘。例如，可挂载/var/lib/docker目录到数据盘，之后再创建/var/lib/docker/timatrix目录，最终创建/data/timatrix软链目录到/var/lib/docker/timatrix目录下。
2. 对于采购了存储节点的场景，需要将存储通过NAS的方式挂载到每台机器的/data/timatrix/storage/volume1。

## 服务器数量评估

对于云智天枢平台，机器数量主要依据具体的业务场景所需的算法资源决定。

常用的计算公式如下（以P40的卡为准）：

1. 2卡gpu机器计算标准

N1 = (P1+P2+P3) / 40

N2 = (P1+P2+P3) / 60

其中：

* N1：人脸布控所需机器数，单位为台；
* N2：车辆布控所需机器数，单位为台；
* P1：人脸路数；
* P2：机动车路数；
* P3：非机动车路数；
* 40：一台2卡机器可以处理40路人脸；
* 60：一台2卡机器可以处理60路车辆。

1. 8卡gpu机器计算标准

N1 = (P1+P2+P3) / 80

N2 = (P1+P2+P3) / 120

其中：

* N1：人脸布控所需机器数，单位为台；
* N2：车辆布控所需机器数，单位为台；
* P1：人脸路数；
* P2：机动车路数；
* P3：非机动车路数；
* 80：一台8卡机器可以处理80路人脸；
* 120：一台8卡机器可以处理120路车辆。

注意：计算说明如下，目前2卡p40和8卡p40，主要瓶颈在cpu，而不是gpu，10路4个cpu，见表3。

*表3 计算说明表*

| 机型 | 人脸结构化 | 人脸布控 | 机动车结构化 | 机动车布控 | 非机动车结构化 | 非机动车布控 |
| --- | --- | --- | --- | --- | --- | --- |
| 2卡p40 | 60 | 40 | 60 | 40 | 60 | 40 |
| 8卡p40 | 120 | 80 | 120 | 80 | 120 | 80 |

1. 云智天枢软件规划

## 操作系统要求

对于部署云智天枢平台的服务器，要求操作系统为CentOS 7.6 (1810版本)，内核要求：Linux version 3.10.0-957.5.1.el7.x86\_64

## 组件拓扑规划

对于组件的拓扑规划，目前有节点数为1、3、5的三种方案可选，请参考《腾讯云智天枢平台实施要求》-非容器化组件拓扑规划。

## 资源消耗型组件和服务的合理分布

一般情况下，对于一个云智天枢集群而言，平台的基础组件（如es、mysql、kafka等）对集群内存、CPU两类资源占用较多，而算法服务对于集群内存、GPU两类资源占用较多。因此，一般的部署策略是：

* 基础组件部署在独立的机器（即组件部署在组件运算节点）上，来分配大量的内存和cpu资源；
* 算法服务作为消耗内存与GPU的资源大户，也会将它们“错位”地部署在别的机器上（即算法服务部署在平台算法运算节点）；
* 为了确保平台业务服务资源消耗不会影响整体集群的调度与容灾，容器平台管理服务也建议单独隔离（即容器平台管理节点）；

1. 云智天枢典型应用场景部署案例

业务场景描述：

用户要求200路人脸布控，人脸小图存储3个月，大图存储7天，结构化数据存储1年。

**资源评估过程：**

各类型的节点的配置和数量计算过程如下，

1. 容器平台管理节点：

使用最低标准，即16核、32G内存、系统盘100G、数据盘300G的机器，需要3台。

1. 平台算法运算节点：

使用最低标准，即28核、128G内存、系统盘100G、数据盘300G、显卡p40\*2的布控机器。

需要的机器数量为：

N = 200 / 40 = 5 台；

1. 组件运算节点：

使用最低标准，即16核、32G内存、系统盘100G、数据盘300G的机器，需要3台，也可以与容器平台管理节点混用。

其中es存储大小单独评估，计算索引所需存储，根据前文所述的公式

S = P \* D \* 1

此处：

* P=200
* D=7

因此索引所需存储为

S = 200 \* 7 \* 1 = 1400G = 1.4T

所以部署es组件的三台机器数据盘至少为1.4T以上，考虑其他预留空间，建议采购2T以上硬盘。

数量为3，可以与也可以与容器平台管理节点或组件运算节点混用。

其中mysql存储大小单独评估，计算结构化存储，根据前文所述的公式

S= ( ( (P1 + P2 + P3) \* M \* 30 \* 10000 ) / 10000000 ) \* 5 + 100

此处：

* P1=200
* P2=0
* P3=0
* M=12

因此，结构化需要的存储空间为

S= ( (200\*12\*30\*10000）/ 10000000 ) \* 5 + 100 = 460 G

所以，评估得到结论是部署mysql组件的机器数据盘至少为460G，考虑其他预留空间，建议可以采购1T以上的硬盘。

机器数量为2，也可以与容器平台管理节点或组件运算节点混用。

1. 存储节点：

存储节点主要需要计算文件存储，节点数量为1，根据前文所述公式

S= ( (P1 + P2 + P3) \* M \* 30 \* 10000 \* (50 + 800) ) / ( 1024 \* 1024 \* 1024 )

此处：

* P1=200
* P2=0
* P3=0
* M=12

因此，存储需要的存储空间为：

S= ( 200\*3\*30\*10000\*50) / (1024\*1024\*1024) ) + (((200)\*7\*10000\*800)/(1024\*1024\*1024))=19T

所以，评估得到的结论是数据盘至少为19T，考虑其他预留空间，建议可以采购最小配置100T的硬盘。

根据以上计算，得出该业务场景下，集群的最小规划方案为3+5+1=9台，高可用规划方案为3+5+2+3+1=14台，详细说明见表4和表5。

*表4 案例评估的最小规划方案*

|  |  |  |
| --- | --- | --- |
| 节点类型 | 数量 | 资源评估备注 |
| 容器平台管理节点/组件运算节点/数据存储节点 | 3 | 16核CPU、32G内存、系统盘100G、数据盘2T，但是由于数据存储节点混用，有两台机器采购的数据盘需要大于3T |
| 平台算法运算节点 | 5 | 28核CPU、128G内存、系统盘100G、数据盘300G、显卡p40\*2） |
| 存储节点 | 1 | 16核CPU、32G内存、系统盘100G、数据盘至少19T |

*表5 案例评估的高可用规划方案*

|  |  |  |
| --- | --- | --- |
| 节点类型 | 数量 | 资源评估备注 |
| 容器平台管理节点 | 3 | 16核CPU、32G内存、系统盘100G、数据盘300G，但是由于数据存储节点混用，有两台机器采购的数据盘需要大于1T |
| 平台算法运算节点 | 5 | 28核CPU、128G内存、系统盘100G、数据盘300G、显卡p40\*2） |
| 组件运算节点 | 3 | 16核CPU、32G内存、系统盘100G、数据盘2T |
| 组件运算节点（mysql独立部署） | 2 | 16核CPU、32G内存、系统盘100G、数据盘1T |
| 存储节点 | 1 | 16核CPU、32G内存、系统盘100G、数据盘至少19T |