

ZENTRIX GATEWAY

企业 AI 产品文档

配置 · 使用 · 结果核对

Zentrix Gateway 部署与服务选型指南

选择托管或私有部署，明确模型与工具所在位置，再约定数据和运维责任。

操作与使用　权限与边界　结果核对

文档版本

2026.10

核对日期

2026 年 10 月 14 日

目录

1. 确定运营方式

2. 三种常见组合

3. 把“不上云”写成可验收方案

4. SaaS 开通与职责

5. 私有部署与实施准备

6. 运维责任

7. 交接与变更

> 2026.10 · 本指南用于方案与交接；实际安装命令、组件版本和资源配置以当次交付的安装器、离线包及对应实施说明为准。
>
> 从下载安装包、环境准备、安装、初始化到验收交付的逐步操作，见《Zentrix Gateway POC 部署手册》（在线阅读）。

> 页面名称按当前源码核对；角色、许可和部署决定实际可用范围。正式操作前，以所在环境显示的配置与状态为准。

1. 确定运营方式

Zentrix Gateway 面向业务系统、开发工具和外部智能体，提供统一模型与工具接入及治理。

产品范围与部署方式分别决定。同一类接入治理需求可以采用 SaaS 托管，也可以部署到企业自己的环境；外部模型和工具的位置再单独决定。

作为中立第三方的接入与治理层，产品支持组合使用不同供应商及企业私有模型。是否支持具体模态、协议和工具参数，应通过真实样例核对，不能由统一入口推定所有上游完全等效。

2. 三种常见组合

组合	软件在哪里运行	模型与工具	应重点确定
SaaS 托管	服务提供方约定的环境	按方案接入外部或专门开放的服务	托管区域、访问方式、数据保留、备份、服务责任
私有部署并接外部服务	企业自有或指定环境	企业允许的公网模型和 SaaS 工具	出网清单、输入范围、第三方授权与费用
私有部署并使用内网依赖	企业允许的内网环境	内网模型、工具、存储及必要依赖	离线物料、内网证书、依赖完整性、更新和恢复方式

“支持 SaaS 服务”既包含产品可按 SaaS 方式提供，也包含连接外部 SaaS 模型和工具。两者可单独成立：私有部署也能接云端工具，SaaS 服务也需要明确其连接范围。



方案示意。数据是否离开企业边界取决于完整组合，不能只看网关装在哪里。

3. 把“不上云”写成可验收方案

不上公有云可以作为私有部署目标，但应把每个依赖落实到允许的网络和资源位置。

对象	需要记录
Web 与 API 入口	业务访问地址与管理访问地址、入口端口与绑定网卡、TLS 证书、访问范围和入口防护
模型推理	上游地址、模型权重位置、推理资源、允许发送的内容
工具与业务系统	API／MCP／外部智能体地址、账号、权限、是否需要外网
数据存储	平台数据库、分析数据库、对象存储的部署方式与位置，调用记录与审计的留存位置
知识与缓存	选用知识库、语义缓存时，其嵌入及匹配模型的位置
认证与通知	SSO、目录同步、SMTP、运维通知渠道的连接对象
安装与更新	安装器、`.zpkg` 离线包、网关节点包与升级包的获取方式；许可证文件的交付方式
证书与基础服务	DNS、时间同步、内部 CA；使用自动签发证书时的 DNS 服务商接口与证书颁发机构
备份与恢复	备份文件转存位置、主密钥保管、异地复制和恢复演练环境

离线环境的边界：

- 安装过程不需要外网，安装所需内容都在安装器与离线包中。离线包与答案文件不包含凭据，凭据在现场安装时经凭据文件或终端录入；许可证在安装完成后由租户管理员导入，并在本地验签。
- 升级包和可选能力包在离线环境中通过“复制获取链接”在可上网的电脑上下载，再上传到“系统更新”页。
- 入口证书的自动签发需要控制面访问证书颁发机构和 DNS 服务商接口。不允许出网时，改用手动上传覆盖入口域名的通配符证书。

验收时核对配置与实际连接，记录允许的目标和业务样例。需要断网验证时，在批准的测试环境中验证指定功能；停用外网后依赖公网的模型、工具和认证不可用属于方案差异，应在验收记录中写明。

使用外部 SaaS 时，发送给上游的提示、文件片段、工具参数和回执可能包含业务信息。应先确定授权范围、数据等级及脱敏要求，再执行调用。

4. SaaS 开通与职责

SaaS 方式由服务提供方运营平台基础设施，包括安装、升级、备份恢复、入口与证书、网关节点和租户开通；企业作为租户管理自己的成员、权限、模型或工具授权及业务使用政策。具体责任以服务协议为准。

1. 确定产品范围、托管区域、租户与默认管理员账号的开通方式，以及租户的结算货币（创建后不能更改）。
2. 明确模型、工具及外部订阅由谁采购和维护，以及租户每月 Token 上限等额度。
3. 由租户管理员导入许可证，配置管理角色、成员、SSO 与可用能力。
4. 对一项业务完成正常调用、权限拒绝、计费归属和问题记录验证。

5. 约定升级通知、事件响应、备份恢复、数据导出和停止服务后的处理。服务运营方可在租户管理中异步导出该租户的全部数据。

事项	服务方与客户需要约定
可用性与支持	服务时段、故障响应、维护窗口与升级通知；升级时控制台切换期间网关调用不受影响，但需约定通知方式
数据	托管位置、调用记录与内容的留存范围、导出和删除流程
费用	软件服务、推理、工具订阅和额外资源分别计费的方式
安全	管理职责、权限审批、凭据轮换与事件通知
退出	导出内容、交接格式、停止访问与数据清理时限

代码中的多租户、租户管理与注册能力说明产品具备相应软件形态，不替代当期 SaaS 套餐、服务区域或服务等级的商务确认。

5. 私有部署与实施准备

5.1 交付形态

私有部署的部署形态在安装时确定，只有以下三个取值，未知取值会使启动和数据库迁移直接失败：

形态	取值	安装方式	说明
单机或多机裸机（Linux）	`bare-metal-systemd`	`zenova install`	服务以 systemd 托管，默认安装到 `/opt/zenova`，可用 `--install-dir` 指定其他目录；多机由控制机按波次经 SSH 安装
单机裸机（macOS）	`bare-metal-launchd`	`zenova install`	服务以 launchd 托管
Kubernetes 多副本	`helm`	Helm chart	按实施说明部署；数据层备份由集群侧负责，能力启停由 chart 配置管理

安装、升级、回滚和备份恢复只由 `zenova` 程序执行，控制台的“系统更新”“备份与恢复”页面与命令行调用同一个执行器。早期的 Docker Compose 离线包不再提供。

Kubernetes 形态的交付范围（包括租户流量入口）以当期实施说明为准，选型前应与实施方确认。

5.2 交付物

交付物	说明
安装器	即 zenova 程序本身，静态发布；版本和平台必须与离线包一致
离线包（.zpkg）	按所选平台、部署形态和能力组装，包含组件与定制站预填的配置；由定制站下载，或由实施方用 zenova package compose 从本地原料组装，两者内容一致
网关节点包（zenova-node.zpkg）	安装远端网关节点时使用
许可证（.lic）	安装完成后由租户管理员在“系统设置 → 全局设置 →「授权许可」页签”导入

安装器启动后先按发布清单逐个校验组件的 sha256 与大小，确认介质完整。

Zentrix 网关版本（安装答案 edition = "zentrix_gateway"）可选的能力如下；托管智能体与文档解析不在 Zentrix 网关版本中提供。

能力	标识	说明
核心	core	必装
知识库	knowledge_base	可选，使用时还需要许可包含该模块
语义缓存	semantic_cache	模型调用的语义缓存；Zentrix 网关版本默认预选
高可用数据层	ha_data	多机且自建数据库时使用

本次未启用的能力保留在本机，之后可在“系统更新”页直接启用。

5.3 主机要求与安装前检查

安装器在写入任何文件之前完成主机检查，每项给出结果；未通过的项目写明原因与处理办法，处理后按回车重新检查。

检查项	要求
CPU 架构	与安装包平台一致（x86_64 或 arm64）
内存	不少于 8 GB
磁盘空间	安装目录（默认 /opt/zenova）所在磁盘：基础 20 GB；自建平台数据库加 30 GB，自建分析数据库加 50 GB，自带对象存储加 10 GB；知识库、语义缓存各加 5 GB。不足时阻断安装（诊断代码 HOST_DISK_LOW），提示可用 --install-dir 安装到空间更大的卷，Linux 主机同时列出可用空间最大的三个挂载点

检查项	要求
系统服务托管	Linux 需要 systemd，macOS 使用 launchd
系统时间	校时服务已同步，时间偏差不超过 30 秒；Linux 通过 `chronyc tracking` 判断
本机网络地址	本机需要可用的非回环网络地址
处理器指令集	在 x86_64 主机上自建分析数据库时，需要 x86-64-v3 指令集（含 AVX2）；虚拟机 CPU 型号偏低时，改为 host 直通或 Haswell 及以上，或将分析数据库改为使用现有服务
端口	业务入口的本机监听端口被占用时报出占用进程并阻断（`HOST_ENTRY_PORT_BUSY`），不自动更换；另行指定的管理入口端口与平台内部服务端口被占用时自动顺延到下一个空闲端口，并在终端说明改用的端口

安装与运维命令都需要 root 权限，请使用 `sudo` 执行。Zentrix 网关版本不包含以容器运行的能力，安装时不要求容器运行时。

需要只做预检时，在命令后加 `--check-only`：安装器完成读包、版本核对、能力选择和主机检查后退出，不写入任何内容。

5.4 数据依赖

平台数据库（PostgreSQL）、分析数据库（ClickHouse）和对象存储逐项选择部署方式。安装后可在“部署配置 → 数据依赖”中查看每项是“由本平台部署”还是“使用现有服务”，以及管理方。

依赖	随安装部署	使用现有服务时的要求
平台数据库	部署在本机（多机时部署在数据层主机）	提供超级用户账号；服务器需提供扩展 `vector`（0.7.0 及以上）、`pg_trgm`、`uuid-ossp`、`pgcrypto`、`btree_gist`；`max_connections` 不低于 400。安装器用该账号创建平台所需的角色与库，管理员密码只在安装过程中使用
分析数据库	部署在本机	版本 26.7 及以上，已定义 `cluster`、`shard`、`replica` 宏，登录账号具备建库权限
对象存储	三选一：“自带对象存储（推荐，生产）”“使用本机目录（单机试用）”	“使用现有的 S3 兼容存储”：提供访问地址、区域、存储桶、寻址方式（路径寻址或虚拟主机寻址）及具有读、写、删权限的访问密钥

补充说明：

- 分析数据库是所有部署形态的必装组件，调用明细、用量统计与策略命中存放其中。单机裸机安装时可选择由本平台部署或使用现有服务；升级前检查新增“分析数据库（ClickHouse）”一项，当前安装未配置分析数据库

时升级被阻止，按提示运行 zenova install --reconfigure 部署或填写现有 ClickHouse 后重新检查。

- 自带对象存储只适用于单机裸机安装，安装器为平台数据、网关报文归档和审计归档分别建桶。
- 使用现有数据库时，连接方式可选明文、加密（校验交给系统信任库）或加密并用指定 CA 证书校验；后两种都校验证书与主机名，区别只在 CA 来源。生产环境建议加密。现有 S3 兼容存储没有单独的加密选项，访问地址以 https:// 开头即加密连接。
- 安装器在录入凭据后、正式安装之前检测现有服务：数据库能否登录、账号权限、扩展与版本、连接数上限、宏定义和建库权限，对象存储的读、写、删。检测只写入并删除临时的探测对象，不写入业务数据。
- 对象存储方式安装后不支持在线修改，需要备份后按新的选择重新安装，再恢复数据。
- 本机备份不包含使用现有服务的平台数据库和 S3 兼容存储，这两部分由数据库管理员和存储管理员按其备份机制保护。密钥服务的密钥目录 /etc/zenova/secret-service/keys 纳入主机级备份，并与备份归档分开保管；换主机恢复时需要该目录。

5.5 单机安装

```
sudo ./zenova install ./<离线包>.zpkg
```

交互式安装依次完成：验证安装包、主机检查、选择能力、确认预填配置、补充未预填的配置（访问地址、各依赖的部署方式、对象存储、运维通知渠道）、录入凭据并检测依赖、发送运维通知测试、确认安装内容。确认之前不写入任何文件，任一步输入 q 退出时本机不留下任何内容。确认后依次安装数据层、控制面（本机的中心网关节点随控制面一并启动）和已选能力，最后执行装后验收。

安装完成时打印控制台地址、用户名 root、初始密码与验收结果。首次登录修改密码之前，可在安装主机执行 sudo zenova root-password show 再次查看初始密码；root 需要重置时执行 sudo zenova root-password set。

非交互安装使用答案文件与凭据文件：

```
sudo ./zenova install ./<离线包>.zpkg --answers answers.toml --secret-file secrets.toml --with core,semantic_cache
```

参数	说明
--install-dir <目录>	安装根，默认 /opt/zenova；程序、数据、版本和解包暂存跟随该目录。后续本机运维命令自动识别，也可显式指定
--log-dir <目录>	独立规范日志根，必须位于安装根外且首次使用时为空；归档与游标账本一起保存。省略时首装跟随安装根，续装沿用已保存配置
--answers <文件>	答案文件；给出后不再交互提问
--secret-file <文件>	凭据文件（权限 0600）；凭据只经该文件或标准输入进入，不写入答案文件
--with <能力>	本次启用的能力，逗号分隔；核心总会包含。不给出时 Zentrix 网关版本启用核心与语义缓存
--reconfigure	保留已有数据，回到选择能力一步，带着上次的答案重新确认

参数	说明
--check-only	只做到主机检查为止，不写入任何内容
--business-entry-port、--admin-entry-port	业务入口、管理入口的本机监听端口；业务入口缺省跟随访问地址的端口（不写端口为 80），访问地址前面有代理或 NAT 时用它指定转发到的本机端口；管理入口缺省与业务入口相同，相同时合并为一个入口
--business-entry-bind、--admin-entry-bind	业务入口、管理入口绑定的本机地址；缺省绑定全部网卡

Linux（systemd）首次安装时，安装器只把业务入口的 HTTP 端口与 443 写入主机防火墙（firewalld 或 iptables）；拆分出的管理入口端口、装后修改过的入口端口以及 macOS 主机的防火墙，由运维方自行放行。除入口端口外，其余端口不需要从主机外部访问，运维方应在主机防火墙或网络边界上只对外放行入口端口。

安装与日志目录应选持久文件系统，Linux 不支持 /tmp、/var/tmp、/home、/root、/run/user、/usr、/etc、/boot；安装盘须允许执行程序。安装根只支持字母、数字、点、下划线、短横线和斜杠。路径配置不提供自动搬迁，一台主机仍只允许一套中心安装。规范日志复用原生按天/500 MiB 分片及 30 天保留策略；系统 journal 和 macOS 有界捕获缓冲仍遵循原生布局。

安装中断或某一步失败时，处理原因后重新执行同一条安装命令，已完成的步骤会跳过。安装完成后的日常配置变更在控制台“部署配置”页或用 zenova config set 完成。

5.6 多机安装

答案文件中写有主机清单（[[hosts]]，每行包含地址、角色和站点）时，在控制机上执行 zenova install 即进入多机安装。主机角色包括控制面、网关节点、平台数据库、分析数据库和仲裁。各主机的安装目录跟随控制机本次的 --install-dir；网关节点主机的安装目录在答案文件中用 gateway_install_dir = "/data/zenova-node" 单独指定（缺省为节点默认目录），不能落在中心安装目录之内。

1. 安装器列出主机计划与缺口，例如某个角色需要的主机数量或站点数量不足；有缺口时只能返回修改或退出。
2. 逐台确认地址、角色与站点，并录入 SSH 登录方式（私钥文件、本机 ssh-agent 或一次性密码）。多台主机可共用一份登录凭据。
3. 输入控制面虚拟地址，必须是单个私网 IP。
4. 写入之前，安装器逐台登录做只读预检（远端以 --check-only 运行），全部通过才开始写入。
5. 按波次安装各主机，进度实时显示。某台主机失败时提示“第 N 波 · 主机 · 原因”；处理后重新执行 zenova install，已完成的波次会跳过。

SSH 凭据只保存在内存中，安装报告只记录地址与结果。

控制机指定的 --install-dir、--log-dir 会同时传给各角色主机的预检与正式安装；修改路径会改变断点身份，不能静默跳过按旧路径完成的主机。网关节点主机不使用中心安装根：控制机指定的 --install-dir 不是默认的 /opt/zenova 时，网关节点也装到该目录，否则装到节点默认目录 /opt/zenova-node；SSH 登台缓存仍位于 /var/lib/zenova/install-stage，需另外保证该文件系统有空间容纳安装器与离线包。

5.7 清空重装与失败续装

目的	命令	对已有数据的处理
升级现有版本	zenova upgrade <升级包>.zpkg	保留数据
重新填写安装配置	zenova install <离线包>.zpkg --reconfigure	保留数据
删除旧部署并重新首装	zenova install <离线包>.zpkg --fresh	清空旧数据、配置与身份后安装新包

清空重装只用于单机裸机安装，且平台数据库、分析数据库和对象存储均为本机自建。存在现有外部数据库或对象存储时，安装器拒绝清空，也不会清理这些资源。网关节点安装、多机与 Kubernetes 部署不使用本流程。早期版本的 Docker 安装（安装根 /opt/lobsterpool）可经本流程清理后改为当前的裸机安装。

1. 使用与新离线包同版本的独立安装器，把安装器、离线包、答案文件与凭据文件放在旧安装根和业务目录之外。
2. 先执行只读预览：sudo ./zenova install ./<离线包>.zpkg --fresh --dry-run。预览列出旧部署的版本、部署形态和清理范围，不停服务、不删除数据。
3. 核对预览并安排停服窗口后执行：sudo ./zenova install ./<离线包>.zpkg --fresh --confirm-data-loss DELETE_EXISTING_DATA。
4. 删除旧数据前，安装器先校验新包组件的摘要与大小、安装器与包的版本和平台、安装答案以及所选能力；未通过时不进入删除阶段。
5. 安装器停止旧服务，清理属于该安装的服务、数据路径和独占资源，再按首装流程安装新包并执行装后验收。

清空重装不自动备份，清理范围包括本机已有的备份。需要保留旧数据时，先备份并将归档移到清理范围之外；恢复只能在同一版本、同一部署形态与平台上进行，旧版本的备份不能恢复到新版本的安装。

中断后的处理：旧部署清理成功、后续安装失败时，处理安装错误后对同一个包执行普通安装命令（不再加 --fresh），从已有安装进度继续。清理阶段失败或中断时，先核对现场残留与失败原因；再次使用 --fresh 并确认删除表示重新清空本机数据。

5.8 网关节点

控制面所在主机本身就是中心网关节点。需要在其他主机就近处理请求时，平台管理员在“网关 → 节点”中添加节点并生成一次性安装命令，在目标主机以 root 运行：

> sudo ./zenova install node --package zenova-node.zpkg --control-plane <控制面回连地址> --token <一次性令牌>

页面要求控制面回连地址以 https:// 开头，节点主机需要能访问该地址；控制面使用自签或内部 CA 时，命令自带该 CA。安装命令有效期 24 小时，过期后在节点详情重新签发。节点与控制面断连期间，请求记录先暂存在节点本地，恢复后补传；控制面不可用期间，网关节点按最近一次下发的配置继续服务。系统升级时，远端网关节点在“网关节点滚动更新”阶段逐台更新。添加节点时可填写“安装目录”（留空则安装到 /opt/zenova-node），节点程序、数据与配置都放在该目录下，命令随之带上 --install-dir <目录>；升级沿用首次安装的目录，不能更换。早期版本使用 /var/lib/zenova-node、/etc/zenova-node 的节点在升级时自动迁移到安装目录下，失败自动恢复原状。

5.9 实施准备与上线顺序

实施前准备：

- 组织与身份：平台管理员、租户管理员与各管理角色，SSO 或本地账号方案。
- 资源与网络：主机规格、网络区段、业务与管理访问地址、入口端口、TLS、DNS 和时间同步。
- 模型与工具：模型来源、推理资源、账号授权及允许连接地址。
- 交付物料：匹配架构、版本和部署形态的安装器、离线包、节点包与许可证。
- 运维：备份转存位置、主密钥与密钥服务密钥目录的保管、恢复目标、运维通知渠道和维护窗口。

平台服务自身与模型推理资源分别估算。安装器的主机检查不包含 GPU 要求；本地模型、嵌入或其他推理组件需要的算力按实际方案另行估算。

上线顺序建议为：基础环境 → 安装与装后验收 → 首次登录与平台账号 → 入口地址与证书 → 网关节点 → 租户与许可证 → 组织和职责 → 模型及工具 → 使用授权与策略 → 普通账号验证 → 告警、备份与恢复演练 → 业务交接。

装后验收通过只证明平台基础服务可用。模型与工具授权、流量规则下发、真实业务和恢复演练仍需逐项完成。

6. 运维责任

事项	私有部署时的负责方	执行入口
安装、升级、回滚、可选能力启停	平台管理员	`zenova install`、`zenova upgrade`、`zenova rollback`；控制台“系统更新”
备份、恢复与恢复演练	平台管理员	`zenova backup`、`zenova restore`；控制台“备份与恢复”
访问地址、入口端口、入口证书	平台管理员，配合网络管理员	控制台“部署配置”“网关 → 入口与证书”；`zenova config set`
网关节点	平台管理员	控制台“网关 → 节点”；`zenova install node`
租户开通与额度	平台管理员	控制台“租户管理”
自建的平台数据库与分析数据库	平台管理员	随平台备份；版本不满足时执行 `zenova infra upgrade`
现有的平台数据库与分析数据库	客户数据库管理员	数据库自身的备份、升级与凭据轮换
现有的 S3 兼容存储	客户存储管理员	存储自身的数据保护与密钥轮换
DNS、防火墙、负载均衡、出口策略	客户网络管理员	按入口端口和出口目标配置
许可证、组织、模型、工具与策略	租户管理员	各租户的管理控制台

主机故障或控制台不可用时，平台管理员在安装主机执行 `sudo zenova status` 查看各服务状态、当前版本与最近一次升级，再按需执行 `sudo zenova restart <服务名>`。

7. 交接与变更

交接内容	应保留的证据
产品与版本	部署形态、版本、已启用能力、许可证与高级模块
部署配置	访问地址、入口端口与绑定网卡、数据依赖部署方式与管理方、最近一次验收结果
外部连接	目标、用途、数据类型、认证、负责人
业务使用	一个正常样例、一个拒绝样例及对应请求标识
运维	告警测试、备份标识与转存位置、实际恢复演练记录
权限	普通成员、管理角色和平台账号的验证结果；root 密码与主密钥的保管人
已知限制	未接入、未启用、未验证和需要外部条件的项目

升级或扩容前，重新核对版本、能力、数据与配置变更、许可、资源和维护窗口。升级前先完成备份并记录备份标识，再按当次实施步骤操作，最后验证关键业务及拒绝路径。

中立第三方定位保留的是供应商选择与统一治理的空间；持续可迁移仍需要维护接口样例、凭据清单、数据导出和替代方案。