CSGHub功能介绍与清单

1. CSGHub平台简介

大模型百花齐放的时代，数据和模型已成为企业和个人用户最核心的数字资产。然而，当前面临的挑战是模型文件和数据集的管理工具分散、管理手段单机化碎片化等问题。这些问题不仅带来了安全运营的巨大风险，也成为企业大模型能力更新迭代的阻力。

CSGHub的使命是为用户提供一个开源开放的、针对大模型原生设计的资产管理平台，支持私有化部署和离线运行。我们的目标是为用户创造一个安全、高效且可信赖的环境，使其能够更好地管理和运用大模型核心资产数据。

CSGHub实现了类似私有化Huggingface的功能，以类比OpenStack Glance管理虚拟机镜像、Harbor管理容器镜像以及Sonatype Nexus管理制品的方式，以开源软件的方式实现对LLM资产的安全、高效管理。

1.1 功能特点

CSGHub平台的功能具有以下几个特点：

- **资产统一管理**： 一站式Hub统一管理模型文件、数据集、大模型应用代码。
- **研发生态兼容**： 提供多种SDK工具，同时支持兼容transform sdk，方便对接用户已有项目。支持HTTPS和SSH协议的Git命令和Web界面操作， 确保不同用户均可方便使用。
- **大模型能力扩展**：原生支持版本化管理、模型格式转化、数据自动预处理、数据集预览等功能。
- **权限与安全**： 支持与企业用户系统集成、支持资产可见范围设置、外内部接口鉴权设计，满足企业安全需求。
- **私有化部署支持**： 无互联网依赖、无云厂商依赖等外部依赖，可一键启动私有化部署。

1.2 技术特点

在技术选型及技术方案上，CSGHub平台具有以下几个特点：

- CSGHub整合了多源Git Server、Git LFS大文件存储协议和对象存储OSS等技术，提供可靠的数据存储层、灵活的基础设施接入层和高兼容的研发工具支持。

- 通过服务化的架构， CSGHub提供CSGHub Server后台服务和CSGHub Web Service的管理界面，普通用户可以快速使用Docker compose或Kubernetes Helm Chart启动服务，实现生产级的资产管理。具备自研能力的用户可利用CSGHub Server进行二次开发，将管理功能集成到外部系统或自定义高级功能。
- CSGHub借助Apache Arrow和DuckDB等优秀开源项目，支持Parquet数据文件格式的预览，便于算法研究人员和爱好者进行本地化数据集管理。
- CSGHub提供直观的Web界面和面向企业组织架构的权限设计，用户可通过Web UI实现版本控制管理、在线浏览和下载，也可以设置数据集和模型文件的可见范围，实现数据安全隔离，还可以对模型和数据集发起话题讨论。

2. 平台功能模块

2.1 资产管理

1. 模型管理
    - 提供模型存储功能，包括机器学习模型和大语言模型
    - 提供新建、删除、修改和查询功能
    - 提供模型版本控制功能，支持模型的迭代和优化
2. 数据集管理
    - 提供数据集存储功能
    - 提供新建、删除、修改和查询功能
    - 提供数据版本控制功能，确保数据变更可追溯
3. 代码仓库管理
    - 提供代码仓库存储功能
    - 提供新建、删除、修改和查询功能
    - 提供版本控制功能
4. 大文件支持
    - 支持超大规模（>100TB) 文件存储
    - 支持大文件的上传、下载
5. Web端在线编辑
    - 支持在Web端预览文本文件
    - 支持在Web端上传、下载文件
    - 支持在Web端在线编辑文件

6. 数据集查看
    - 支持在Web端预览查看Parquet格式的数据集，便于快速检查和分析
7. 资产元数据管理
    - 为各项资产提供详尽的分类标签，便于快速检索和管理
    - 支持模型的任务、语言、框架、License标签管理
    - 支持数据集的任务、语言、License标签管理
    - 支持代码的License标签管理

2.2 AI应用

1. 应用空间
    - 用户可以在应用空间内快速展示模型的能力，搭建并测试应用原型，提升开发效率
    - 支持用户进行应用的新建、删除、修改和查询操作
    - 集成了Gradio和Streamlit等流行的应用SDK，用户可以轻松地搭建交互式应用
2. 一键启动模型推理服务
    - 用户可以一键启动模型的推理服务，无需复杂的配置，简化部署流程
    - 自动适配优化运行环境，确保模型推理服务的稳定性和高效性
    - 提供的推理服务可以轻松集成到现有的应用中，加速应用的开发和上线
    - 支持vllm，TGI等主流推理框架
3. 一键启动模型微调训练
    - 使用户能够一键开启针对特定数据集的模型微调训练任务
    - 支持不同的训练参数自定义设置，为用户提供灵活的微调选项
    - 集成训练监控功能，用户可以实时跟踪训练进度和性能指标
4. 模型推理优化
    - 针对特定模型，提供模型推理调优服务
    - 针对企业的业务场景，提升模型在生产环境中的表现
    - 根据不同企业和模型的具体需求，提供定制化的推理优化解决方案
5. 数据格式转化
    - 支持常见的数据格式转化，如将CSV、JSON等格式转换为模型训练或推理所需的数据格式
    - 提供易于使用的转换工具，用户可以快速完成数据格式的转换工作
    - 提供数据处理工具，完成从数据采集、清洗、标注到入库的一整套流程
6. 资产Copilot

- 提供资产管理助手，辅助用户更便捷的使用CSGHub

7. 私有化模型版资产Copilot
   - 提供资产管理助手的私有化部署

2.3 多源数据同步

1. 查看OpenCSG传神社区源数据
   - 启动CSGHub之后可实时浏览查看传神社区的模型及数据集，及时获取最新或者最热门的模型和数据集
   - 可实时感知用户同步的模型数据源是否有更新
2. 模型/数据集推荐
   - 可根据企业实际的业务场景推荐模型
   - 可根据微调模型的特点推荐模型和数据集
3. 同步OpenCSG传神社区模型
   - 支持通过同步服务下载传神社区的模型（不同版本资源限制不同）
4. 同步OpenCSG传神社区数据集
   - 支持通过同步服务下载传神社区的数据集（不同版本资源限制不同）

2.4 安全合规

1. 自定义资产元数据
   - 允许用户自定义资产标签，为用户提供更灵活的资产标记和分类方法
   - 通过自定义标签，增强资产检索的准确性和便捷性
2. 模型和License合规性溯源与验证
   - 对模型使用的许可证（License）进行合规性检查，确保所有使用的模型和库符合所需的授权条件
   - 提供溯源功能，以确保模型及其组件的使用完全符合法律和规范要求
3. 数据完整性校验机制
   - 采用高效的数据完整性校验算法，确保数据在传输和存储过程中保持未被篡改。
   - 及时检测并纠正可能出现的数据错误，确保数据的准确性
4. 高可用
   - 构建高可用的系统架构，确保服务在各种情况下均能稳定运行。
   - 通过冗余备份机制，提高系统容错能力，避免单点故障导致的服务中断
   - 实现负载均衡和动态资源调度，确保高并发情况下的服务响应速度和稳定性

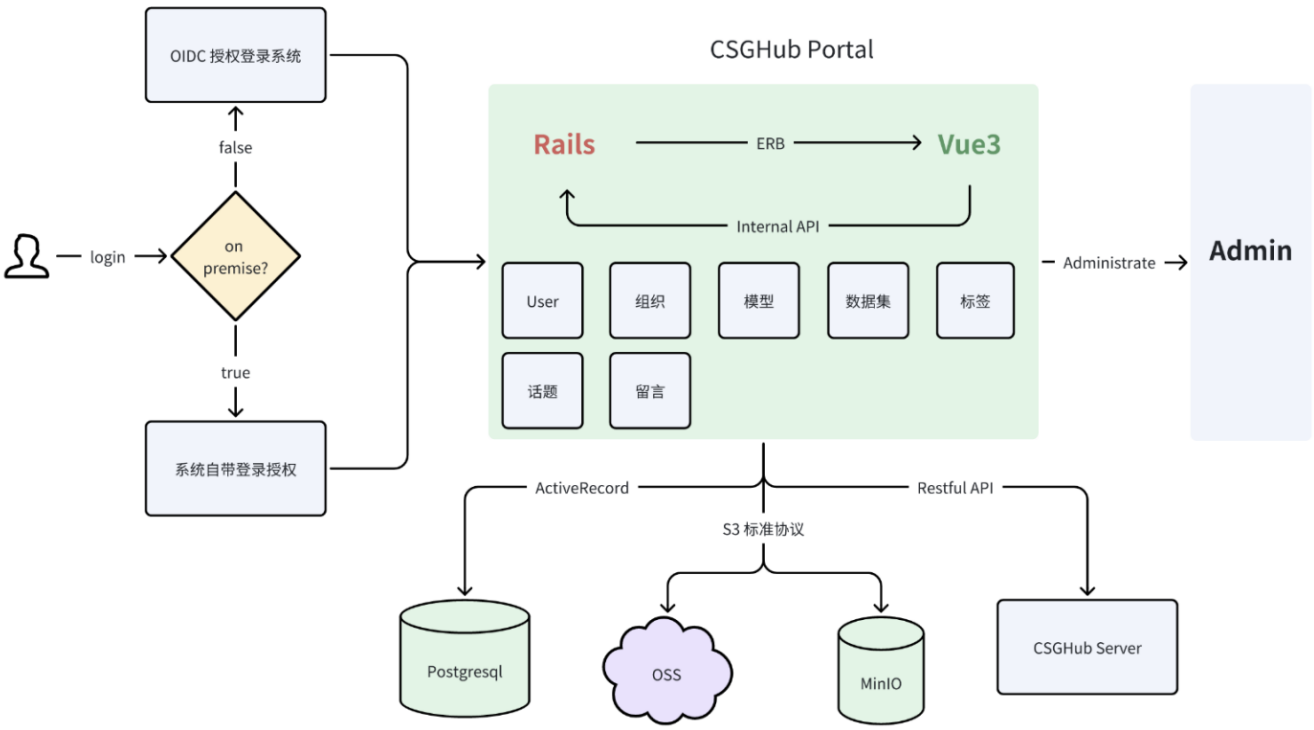
5. 灾难恢复
    - 设计灾难恢复计划，确保在出现系统故障时可以快速恢复服务
    - 通过定期备份和快照技术保护关键数据
    - 制定详细的灾难恢复策略和步骤，降低灾难发生对业务连续性的影响

3. 平台技术架构

CSGHub由Portal和Server两部分服务组成，CSGHub Portal是由vue和Ruby实现，而CSGHub Server则是由Golang实现的一个高性能后端。

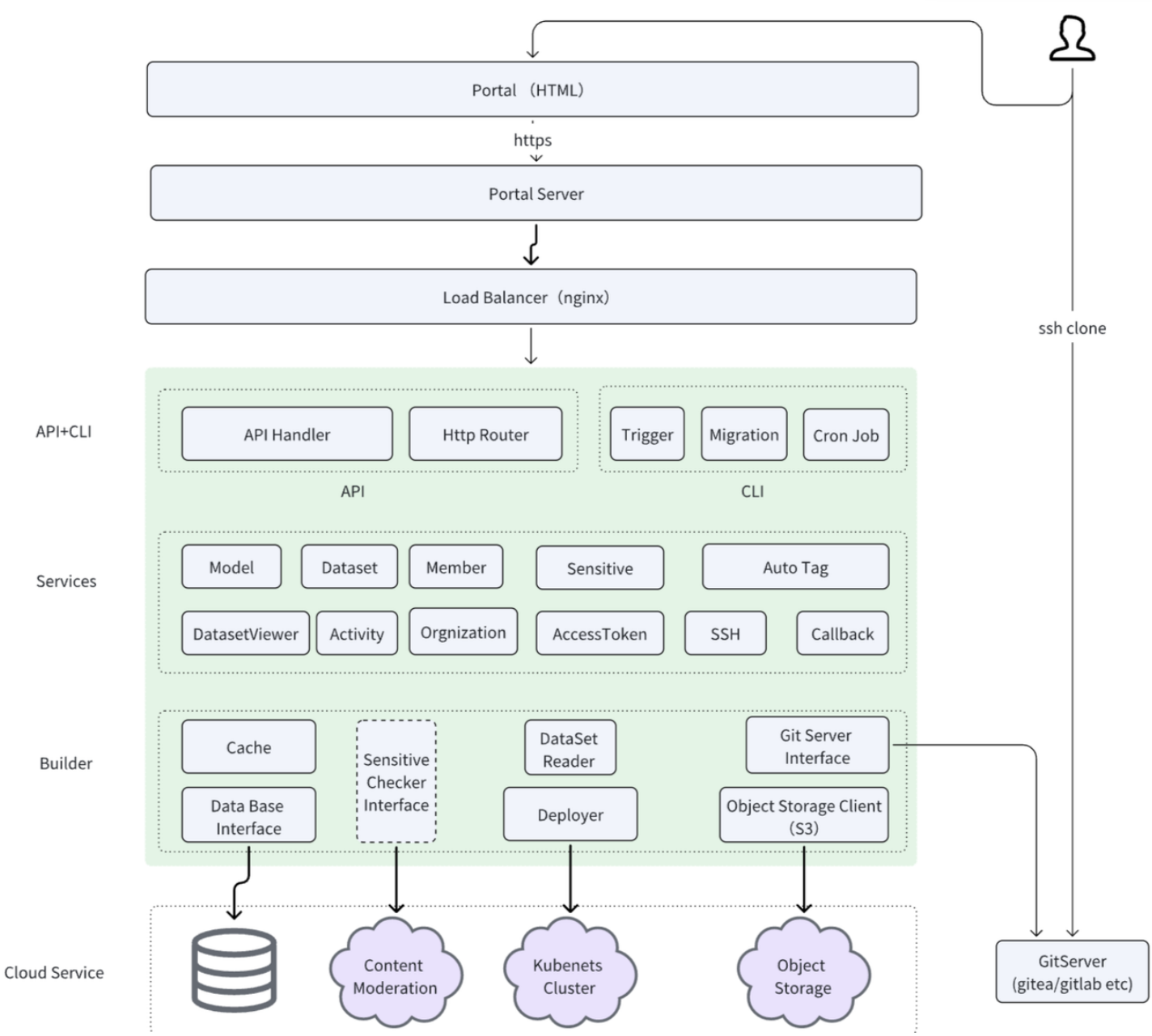
3.1 前端架构

Portal部分的架构设计如下：



3.2 后端架构

Server部分整体架构如下所示：



根据用户实际情况，具有极高的扩展性和灵活性：

- 支持不同的后端Git Server，如gitea，gitlab，local git server等
- 可选择使用本地或第三方兼容S3协议的任意云存储服务，作为LFS存储系统的后端
- 根据用户需要，按需开启内容审核，选择任意第三方内容审核服务。默认关闭，当前支持Aliyun内容安全服务。
- 支持多种关系型数据库，如SQLite，Postgresql， MySQL等。

注意：

当前CSGHub Server自身仅仅支持以标准REST方式对外提供服务，需要借助于CSGHub Portal提供GUI。

4. 技术性能指标

4.1 高效存储与处理能力

- 数据管理平台能够高效地存储和检索大规模数据集，支持高速数据读写操作，确保数据的实时性和准确性。
- 支持不小于100TB规模的单个数据集存储。
- 支持不小于100TB规模的单个模型存储

4.2 模型/数据的兼容性

- 平台支持多种数据格式的导入和导出，具备良好的数据兼容性，能够满足不同数据来源和格式的需求。
- 平台支持多种模型格式转化，以满足不同训练推理框架的要求。
- 平台支持多种机器学习算法和深度学习框架，用户可以根据需求选择合适的算法进行模型训练，实现最佳的训练效果。

4.3 安全可靠性

- 采用先进的数据加密和备份技术，确保数据的安全性和可靠性，防止数据泄露和丢失。在安全性方面提供数据加密和访问控制机制。
- 支持功能、数据、模型级别的权限控制
- 在安全性方面提供用户认证和访问控制机制

5. 产品优势

5.1 特点

- 开源与商业友好的完美结合：CSGHub采用阿帕奇许可证（Apache 2），这意味着用户可以在遵守许可证条款的前提下，自由地使用、修改和分发项目的源代码。这种灵活性使得CSGHub能够轻松适应各种商业场景，同时保证了项目的持续发展与完善。
- 强大的社区支持与国内贡献：CSGHub项目汇聚了众多国内开发者，其中，国内贡献者占比超过80%，显示出项目在国内的广泛影响力和深厚的技术积累。特别值得一提的是，我们公司作为项目的主导者，贡献者占比超过60%，为项目的稳定性和持续发展提供了有力保障。
- 商业应用实证：CSGHub已经成功助力多家商业客户实现业务落地，这充分证明了项目的实用性和可靠性。通过实际应用的检验，CSGHub展现出了其卓越的性能和灵活的定制性，能够满足不同企业的个性化需求。

5.2 价值

- CSGHub项目不仅具备技术上的先进性，更在商业应用上展现了巨大的潜力。通过引入CSGHub，企业可以快速搭建起稳定、高效的软件平台，从而降低成本、提升效率，并在激烈的市场竞争中占据有利地位。

6. 商业版与社区版功能对比

CSGHub的社区版地址：https://github.com/OpenCSGs/csghub

针对企业的使用场景，我们为企业提供商业版本，除了社区版的基础功能之外，在AI应用、多源数据同步、安全合规的模块都提供了更多的增强功能。另外针对企业版的客户，我们提供界面定制服务以及企业级细粒度权限控制定制服务，同时我们还为企业版的客户提供专属技术支持服务。

OpenCSG
开源开放 智慧传神

CSGHub 商业版 VS 社区版

产品模块	功能点	CSGHub商业版	CSGHub社区版
资产管理	模型管理	✓	✓
	数据集管理	✓	✓
	代码仓库管理	✓	✓
	大文件支持	✓	✓
	Web端在线编辑	✓	✓
	数据集查看	✓	✓
	资产元数据管理机制	✓	✓
AI应用	应用空间	✓	✓
	一键启动模型推理服务	✓	✓
	一键启动模型微调训练	✓	✓
	模型推理优化	✓	×
	数据格式转化	✓	✓
	资产Copilot	✓	✓
	私有化模型版资产Copilot	✓	×
多源数据同步	查看OpenCSG传神社区源数据	✓	✓
	模型/数据集推荐	✓	×
	同步OpenCSG传神社区模型	✓	数量、速度受限
	同步OpenCSG传神社区数据集	✓	数量、速度受限
安全合规	自定义资产元数据	✓	×
	模型和License合规性溯源与验证	✓	×
	数据完整性校验机制	✓	×
	高可用	✓	×
	灾难恢复	✓	×
定制功能	企业级细粒度权限控制	✓	×
	界面定制	✓	×
支持服务	技术支持服务	商业支持	社区支持
	获得技术支持服务的通道	OpenCSG官方工单系统	GitHub或者OpenCSG社区
	产品故障的技术支持响应级别	工单系统支持，提供7x24在线服务，针对生产运行中遇到的问题，提供最高30分钟内响应级别。	无
	咨询、规划、实施、主动式巡检、故障排查、重要时期保障	商业专家支持服务（远程+现场）	无

7. 竞争力对比分析

CSGHub竞争力对比分析

	Huggingface	魔搭社区	CSGHub	GiteaAI	始智社区	Ollama
模型社区	✓	✓	✓	✓	✓	✗
国内高速访问	✗	✓	✓	✓	✓	✓
在线微调/推理	✓	✓	✓	✓	✓	✗
研发 SDK	✓	✓	✓	✗	✗	✗
平台开源	✗	✗	✓	✗	✗	✓
离线部署	✗	✗	✓	✗	✗	✓
上游资源联动	✓	✓	✓	✗	✗	✗
与MaaS集成	✗	✗	✓	✗	✗	✓
离线微调/推理	✗	✗	✓	✗	✗	✓
合规与溯源	✗	✗	✓	✗	✗	✗