基于主动学习的标注云平台

项目概要设计

目录

Contents

功能

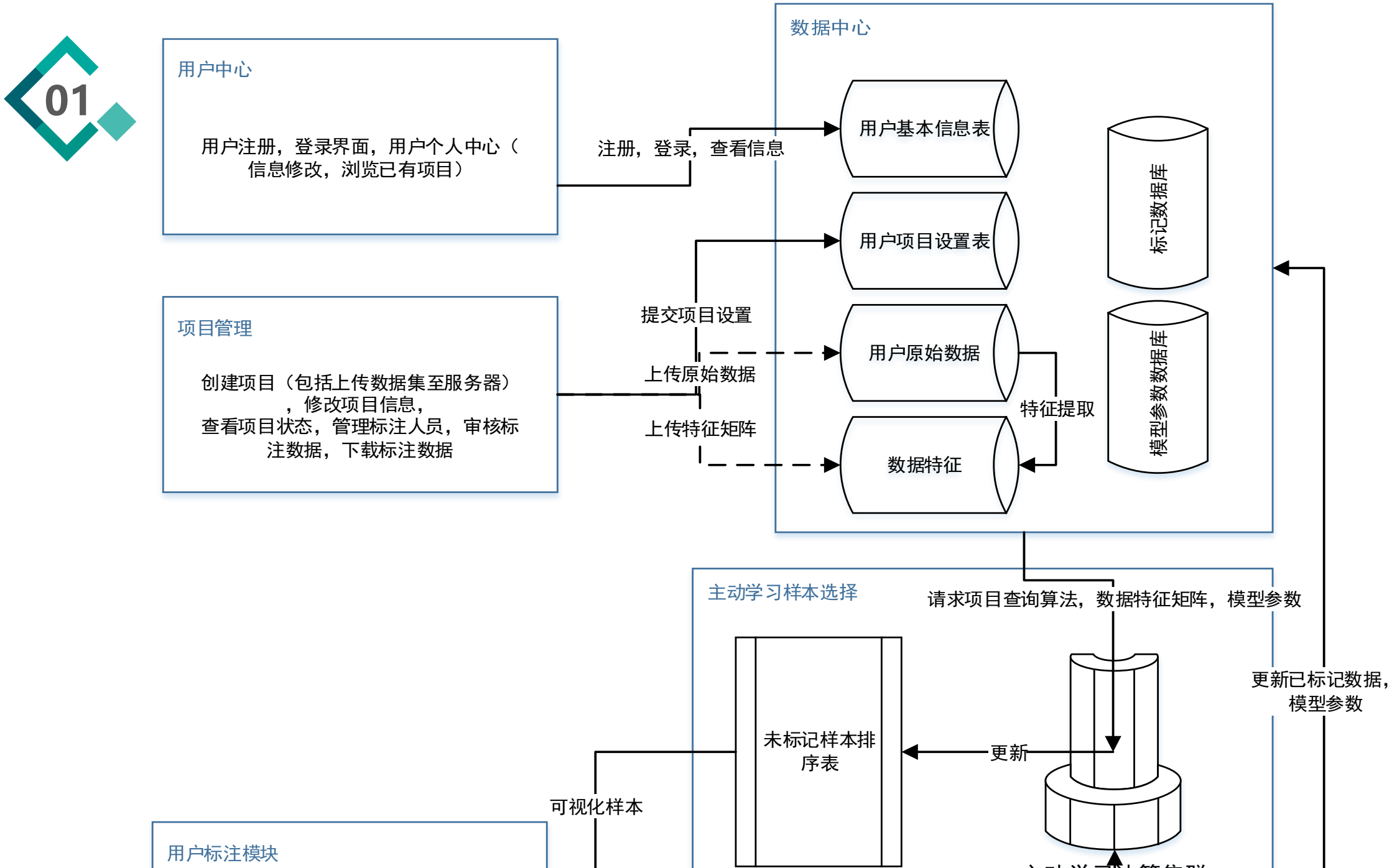
1. 前端：

1.1. 用户注册/登录
1.2. 用户创建项目
1.3. 更改项目设置
1.4. 项目持有者查看项目状态
1.5. 项目持有者管理标注者
1.6. 可视化多种类型的数据
1.7. 支持多样化的监督信息
1.8. 下载标记
1.9. 系统管理员界面

2. 后端

2.1. 数据的可靠存储
2.2. 数据的高效读写与传输
2.3. 主动学习算法的高效计算
2.4. 特征提取
2.5. 模型训练
2.6. 以多种流行的存储格式保存标注信息
2.7. 上传/下载

01


用户中心
用户注册，登录界面，用户个人中心（
信息修改，浏览已有项目）
项目管理
创建项目（包括上传数据集至服务器）
，修改项目信息，
查看项目状态，管理标注人员，审核标
注数据，下载标注数据
注册，登录，查看信息
提交项目设置
上传原始数据
上传特征矩阵
数据中心
用户基本信息表
用户项目设置表
用户原始数据
数据特征
特征提取
标记数据库
模型参数数据库
请求项目查询算法，数据特征矩阵，模型参数
主动学习样本选择
未标记样本排
序表
更新
可视化样本
更新已标记数据，
模型参数
用户标注模块

需求分析

• 用户能够实时地标注数据	维护一个样本排序表
• 后端能够接受大量的异步请求（计算，读写）	分布式系统+消息队列
• 保障用户上传数据的安全性	严格记录操作日志(用于恢复)+灾难备份
• 计算节点需要大量数据	将数据库拷贝(同步)至每一个计算节点
• 特征/标记矩阵的存储与传输与同步	缓存+SQL持久同步存储
• 深度模型的存储与传输	GPU服务器负责训练与预测，不传输模型

需要用到的技术

- ✓ 普通与深度模型的训练 (sklearn + pytorch)
- ✓ 主动学习查询算法 (ALiPy)
- ✓ 特征提取
- ☐ Web框架
- ☐ 消息队列
- ☐ Python分布式框架
- ☐ 操作关系型数据库与非关系型数据库
- ☐ 数据库同步

02 调研

amazon mechanical turk

Pick the best category

Requester: Zeon Reward: $0.00 per task Tasks available: 0 Duration: 1 Hours

Qualifications Required: None

Instructions

View full instructions

View tool guide

Read the task carefully and inspect the image.

Choose the appropriate label that best suits the image.

Choose the correct category

Image will display here

Select an option

1. Cat
2. Dog
3. Bird
4. None of the Above

Zoom in Zoom out Move Fit image

Submit

02 调研

消息队列 • 后端能够接受大量的异步请求（计算，读写）

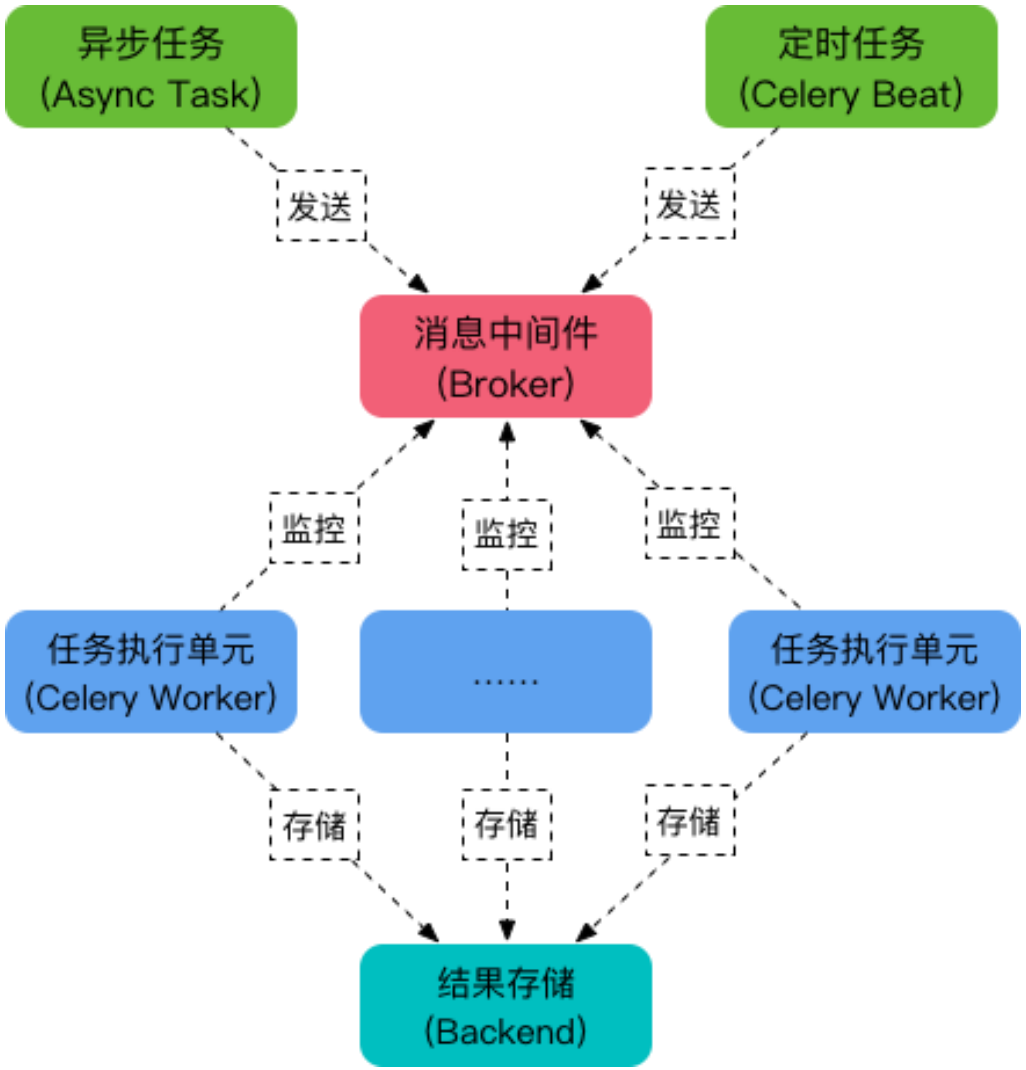
保证：单次送达，有序，
日志记录，性能分析

现有的常用Python消息队列有： rebbitMQ, Redis

02 调研

Celery: Distributed Task Queue

分布式框架



- 后端能够接受大量的异步请求（计算，读写）

调研

- 操作关系型数据库与非关系型数据库

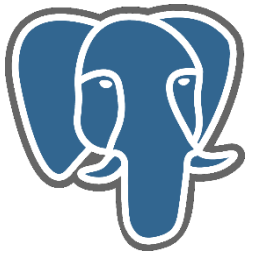
非关系型数据库的优势：

1. 不需要经过SQL层的解析，所以性能非常高，支持高并发操作。
2. 基于键值对，数据之间没有耦合性，所以非常容易水平扩展。

Redis

关系型数据库的优势：

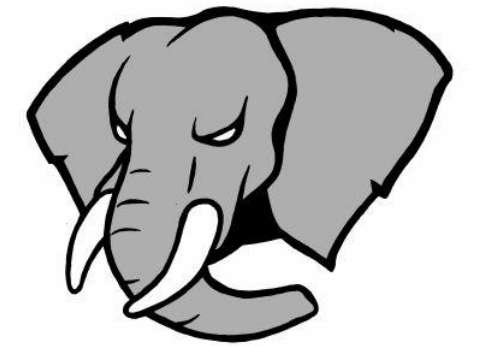
1. 可以用SQL语句方便的在一个表以及多个表之间做非常复杂的数据查询。
2. 事务支持使得对于安全性能很高的数据访问要求得以实现。



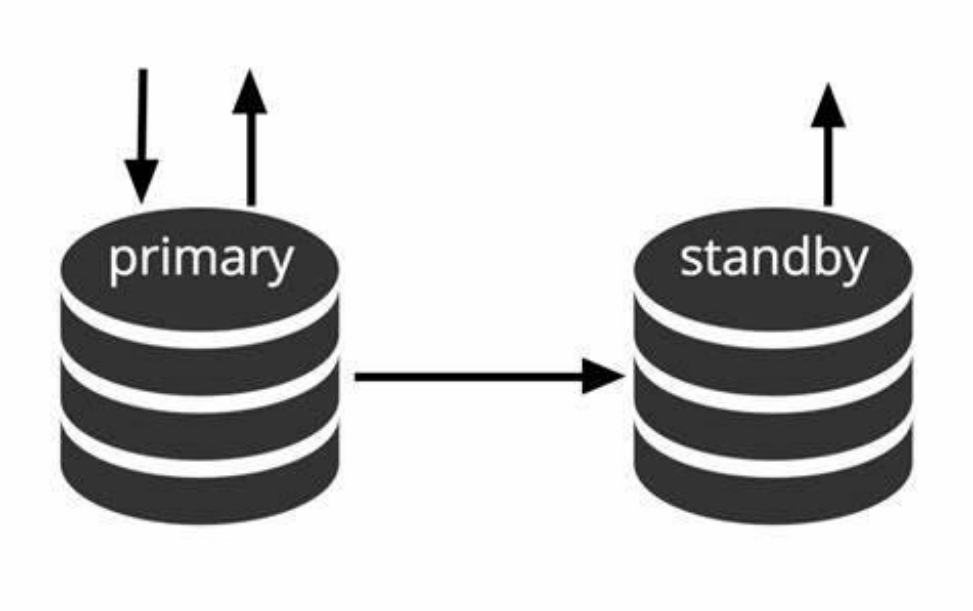
PostgreSQL

02 调研

- 数据库同步
- 特征/标记矩阵的存储与传输与同步



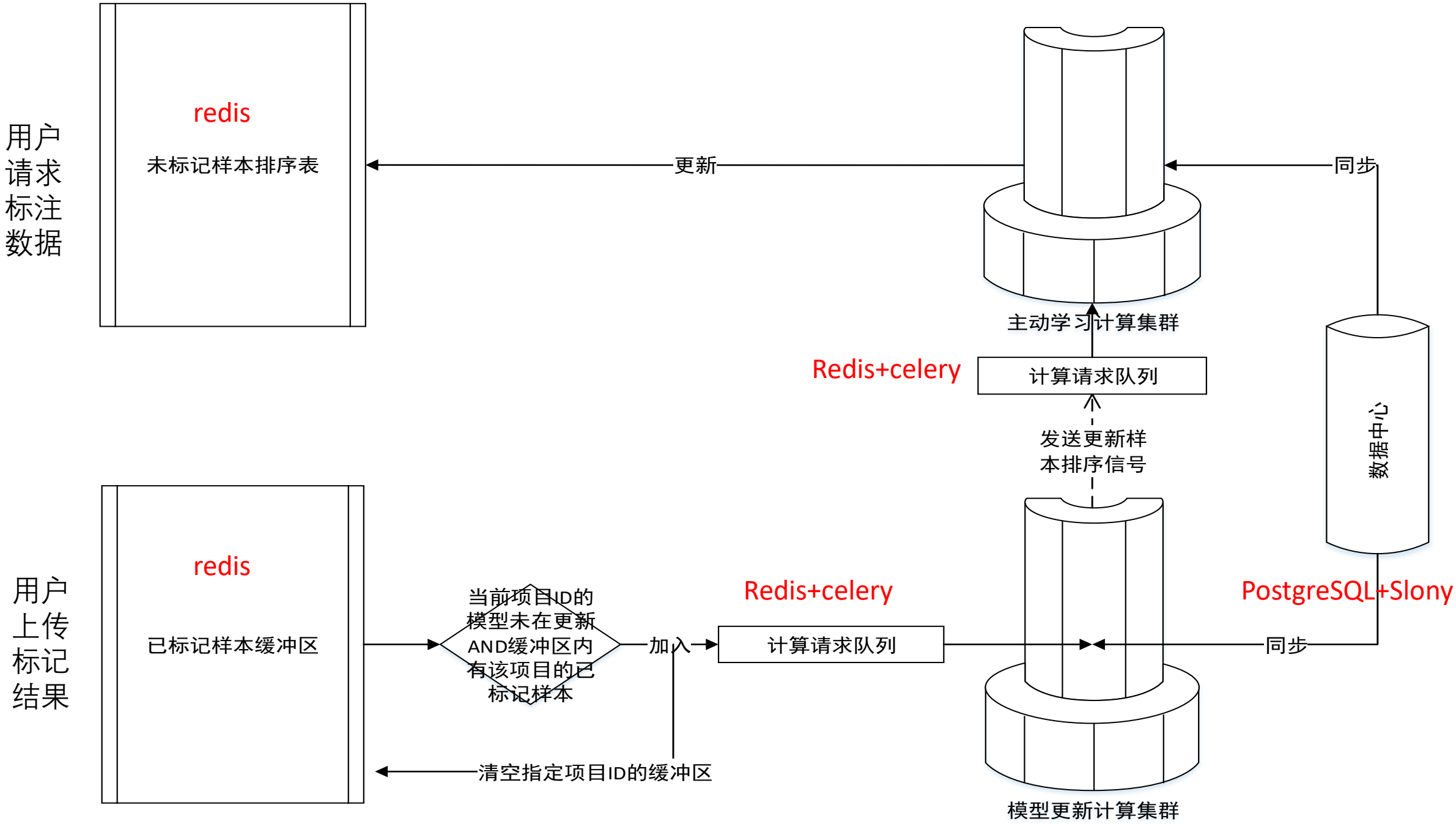
Slony-I is a "master to multiple slaves" replication system for PostgreSQL supporting cascading.



后端设计

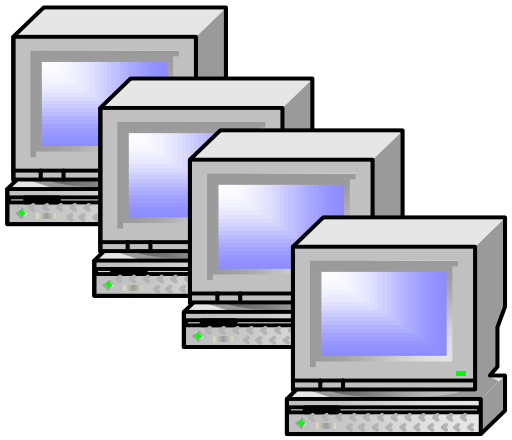
设计原则

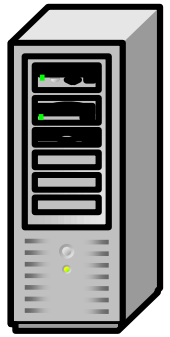
- 减少数据传输
- 分布式处理请求
- 减少中间操作


用户请求标注数据
redis
未标记样本排序表
更新
同步
主动学习计算集群
Redis+celery
计算请求队列
发送更新样本排序信号
数据中心
用户上传标记结果
redis
已标记样本缓冲区
当前项目ID的模型未在更新AND缓冲区内有该项目的已标记样本
加入
Redis+celery
计算请求队列
同步
PostgreSQL+Slony
清空指定项目ID的缓冲区
模型更新计算集群

03 模型训练集群

普通模型的重新训练，预测





深度模型的(增量)训练，预测，本地保存/读取模型

CPU服务器

GPU服务器

任务分配

Celery

数据库设计

	读需求	写需求	安全需求	解决方案
用户基本信息	中	中	高	关系型数据库
项目设置	高	低	高	关系型数据库
原始数据集(包括上传的特征矩阵)	低	低	高	文件存储
项目计算状态	高	高	低	NoSQL
单个数据状态	高	高	高	关系型数据库(需要复杂检索操作)
数据特征矩阵	高	中	低	关系型数据库并同步到从节点
深度模型参数	高	高	低	文件存储
下标管理	高	高	低	NoSQL
标记数据	高	高	高	关系型数据库(需要复杂SQL查询)并同步到从节点