数据标注平台

目录

表 1- 1 文档信息

文档名称	S2C 技术路线及实现方案		
负责人	技术经理	文档版本编号	1.0.6
项目阶段	完工	文档版本日期	2023/11/27
起草人	技术组长	起草日期	2023/10/30

表 1- 2 修改历史纪录

日期	版本	变更说明	作者
2023/10/30	1.0.0	初稿	技术组长
2023/11/03	1.0.1	系统架构完善	技术经理
2023/11/07	1.0.2	ER 图完善	技术组长
2023/11/14	1.0.3	关键技术改动	技术经理
2023/11/18	1.0.4	图片微调	项目经理
2023/11/22	1.0.5	文字细节改动	技术组长
2023/11/27	1.0.6	格式改动	技术组长

1. 技术开发框架

1.1. 架构图

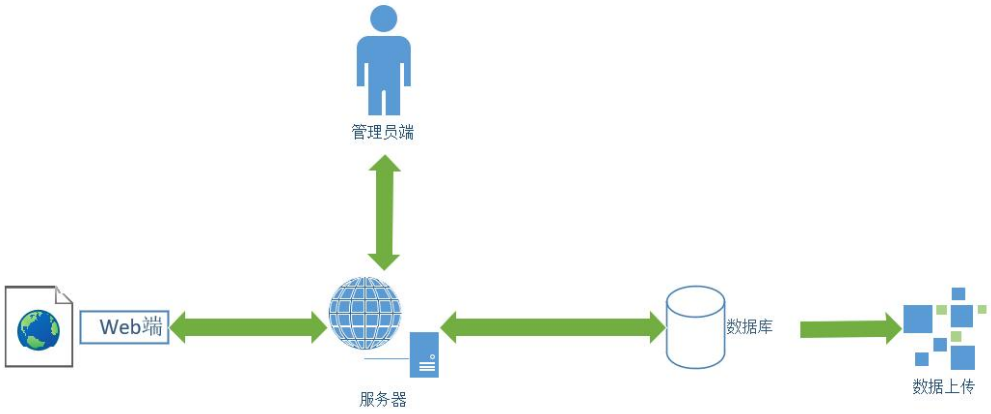

图 1- 1 系统总体框架图

2. 关键开发技术

2.1. 数据标注

本系统为 Web 端可视化数据标注页面，能为计算机视觉项目、自然语言处理研究还是数据科学工作提供卓越的数据标注体验和价值。

在此基础上我们的数据采集模块具有以下优点：

- 直观易用：我们的界面设计简洁明了，让用户可以轻松快速地进行标注工作，无需复杂的培训。
- 多样化的标注工具：我们提供自定义四边形标注框，以满足不同类型数据的标注需求。
- 高度可定制：您可以根据项目需求自定义标注标签，以确保数据标注的一致性和准确性。
- 预标注：我们将自动化工具与标注页面集成，以加速标注过程，例如预设标签或文本识别。
- 数据管理和版本控制：我们提供强大的数据管理和版本控制功能，确保数据的安全性和可追溯性。
- 安全和隐私：我们采取严格的安全措施，保护您的数据安全和隐私，符合数据保护法规。

2.2. 数据审核--对标注结果进行二次审核

我们的 Web 可视化数据标注工具不仅提供了强大的标注功能，还包括了一整套标注审核流程，以确保标注结果的准确性和一致性。

- 初步审核和修正： 用户可以轻松地对标注结果进行初步审核，检查是否存在明显的错误或不一致之处。如果发现问题，可以立即进行修正，确保标注质量。
- 二次审核和验证： 我们支持多级审核，允许第二个审核者对标注结果进行验证。这有助于排除主观性差异和进一步提高标注的准确性。
- 错误反馈和修正历史： 工具记录了每次审核和修正的历史，使用户可以追溯到之前的操作。这有助于分析问题并追踪改进的进展。
- 统计和分析： 我们提供了强大的统计和分析功能，帮助用户了解标注结果

的质量，包括准确性、一致性和进度。这有助于优化标注流程和分配资源。

- 自定义审核规则： 用户可以根据项目需求自定义审核规则和标准，确保审核过程与具体任务一致，并满足特定的质量标准。
- 协作和通信工具： 我们的工具支持团队内的协作和沟通，用户可以轻松地共享反馈、讨论问题，并协作解决标注质量问题。
- 版本控制： 标注审核过程中的每个版本都得到了记录和保存，确保您可以随时查看和比较不同版本的标注结果。

2.3. 数据发布--生成当前的数据快照

数据发布是数据标注工作流程的重要环节，它涉及将已经通过审核的标注数据集进行有效管理、存储和分享的过程。以下是我们的 Web 可视化数据标注工具在数据发布方面的关键特点和步骤：

- 唯一版本号： 每次发布都会生成一个唯一的版本号，以确保数据集的版本管理清晰明了。这有助于跟踪不同发布版本之间的变化和进展。
- 数据快照： 在发布时，会为数据集生成当前的数据快照，这意味着发布的数据集是一个静态快照，不可进行修改。这有助于保持数据的完整性和可追溯性。
- 权限控制： 工具允许具备权限的用户进行数据发布，并且可以对数据集的可见性和访问权限进行精确控制。这确保了只有授权用户可以查看和导出数据集。
- 导出选项： 具备权限的用户可以选择将已发布的数据集导出到不同格式，以满足不同应用的需求。支持常见的数据格式，如 CSV、JSON、XML 等。
- 元数据和文档： 在发布时，用户可以添加元数据和文档，以提供关于数据集的详细信息，包括数据来源、标注规范、数据结构等。这有助于其他用户了解数据集的背景和用途。
- 历史记录和审计： 所有的数据发布操作都得到记录和审计，以确保数据集的安全性和可追溯性。这有助于跟踪数据集的使用历史和管理发布权限。

2.4. 项目管理--不同项目之间数据相互隔离

项目管理是数据标注工作流程中的基本组成部分，它有助于有效组织和管理不同数据标注任务，并确保数据的安全性和隔离。

以下是我们的 Web 可视化数据标注工具在项目管理方面的关键特点和功能：

- 项目隔离： 所有的标注数据都按照项目进行隔离，不同项目之间的数据相互独立。这确保了数据的安全性和隐私，同时使不同项目的数据管理更加清晰。
- 权限控制： 用户只有拥有项目的权限，才能够对项目下属的数据进行标注、审核以及发布。这有助于确保只有授权用户可以访问和处理特定项目的数据。
- 多项目支持： 工具支持创建和管理多个项目，用户可以根据不同的任务和需求组织项目，同时轻松切换和管理这些项目。
- 项目元数据： 用户可以为每个项目添加元数据，包括项目名称、描述、负责人等信息，以便更好地理解项目的背景和目的。
- 任务分配和跟踪： 项目管理界面允许项目负责人分配标注任务给团队成员，并跟踪任务的进展和完成情况。这有助于分工合作和任务管理。
- 数据共享： 具备权限的用户可以在项目内部共享数据，以便多个团队成员协同标注和审核。这加速了标注工作流程，提高了工作效率。
- 项目历史记录： 工具记录了项目的历史操作和活动，用户可以随时查看项目的变更历史，以便了解项目的演进。

2.5. 权限管理

- 权限管理是确保系统安全性和数据保护的关键组成部分。我们的 Web 可视化数据标注工具提供了强大的角色、用户和权限管理功能，以满足不同用户需求，并确保系统操作的合规性。以下是关于权限管理的关键特点和功能：
- 角色分配： 管理人员可以创建不同的角色，每个角色可以具有不同的权限和访问级别。例如，管理员、标注员、审核员等不同角色可以访问不同的功能和数据。
- 用户管理： 管理人员可以添加、删除和编辑用户账户，并将用户分配到适当的角色。这有助于灵活管理系统的用户。
- 项目权限： 用户可以被分配到一个或多个项目，只有当用户被分配到项目时，才能看到该项目以及与之相关的数据。这确保了数据的隔离和安全性。
- 菜单和按钮权限控制： 管理员可以配置对系统的菜单和按钮进行权限控制。这意味着只有具备相应权限的用户才能访问特定功能，从而确保系统操作的合规性。
- 自定义权限规则： 工具支持自定义权限规则，管理人员可以根据实际需求创建自定义的权限策略，以满足特定项目或团队的要求。
- 审计日志： 系统记录了用户操作的审计日志，包括登录、权限更改、数据

访问等，以便审计和追踪系统操作的历史。

- 密码安全： 用户密码得到加密存储，并且支持密码策略，以确保用户账户的安全性。

3. 数据可视化

3.1. 前端交互技术

整个前端使用 HTML5 编写完成。

HTML5 是一种用于构建现代 Web 应用程序的标准，它引入了许多重要的特性和改进，使 Web 开发变得更加强大和灵活。

以下是 HTML5 的一些主要优点：

- Canvas 绘图： HTML5 引入了<canvas>元素，允许开发人员使用 JavaScript 绘制图形、图像和动画，为游戏和数据可视化等领域提供了强大的工具。
- 本地存储： HTML5 提供了本地存储选项，如 Web Storage 和 IndexedDB，使 Web 应用能够在本地存储数据，提高了离线应用的能力。
- 语义化标记： HTML5 引入了许多新的语义化元素，如<header>、<footer>、<nav>、<article>等，使网页的结构更加清晰和可读，有助于搜索引擎优化（SEO）。
- 表单控件增强： HTML5 改进了表单控件，引入了新的输入类型（如<input type="date">和<input type="email">）以及表单验证功能，提高了用户体验和数据验证的效率。
- 跨平台兼容性： HTML5 标准被广泛支持，几乎所有现代的 Web 浏览器都支持 HTML5，使开发人员能够创建跨平台的 Web 应用。
- 响应式 Web 设计： HTML5 与 CSS3 结合使用，使响应式 Web 设计变得更加容易，使网页能够自动适应不同设备和屏幕大小，提供更好的用户体验。

3.2. 数据可视化

数据可视化是利用计算机图形学的图像处理技术，将数据转换成图形或图像在屏幕上显示出来，并进行交互处理的理论、方法和技术。

3.2.1. 模块介绍

- 数据标注： 提供信息抽取标、注文本分类标注、图像文本标注、图片分类标注功能

3.2.2. 界面展示

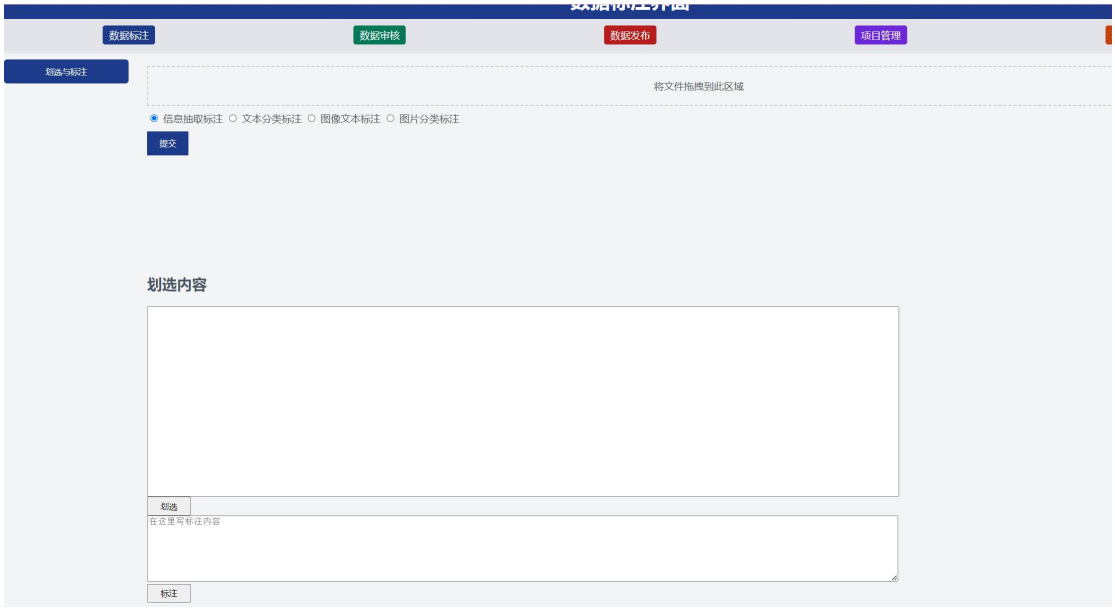

图 3- 1 信息抽取标注

图 3- 2 文本分类标注

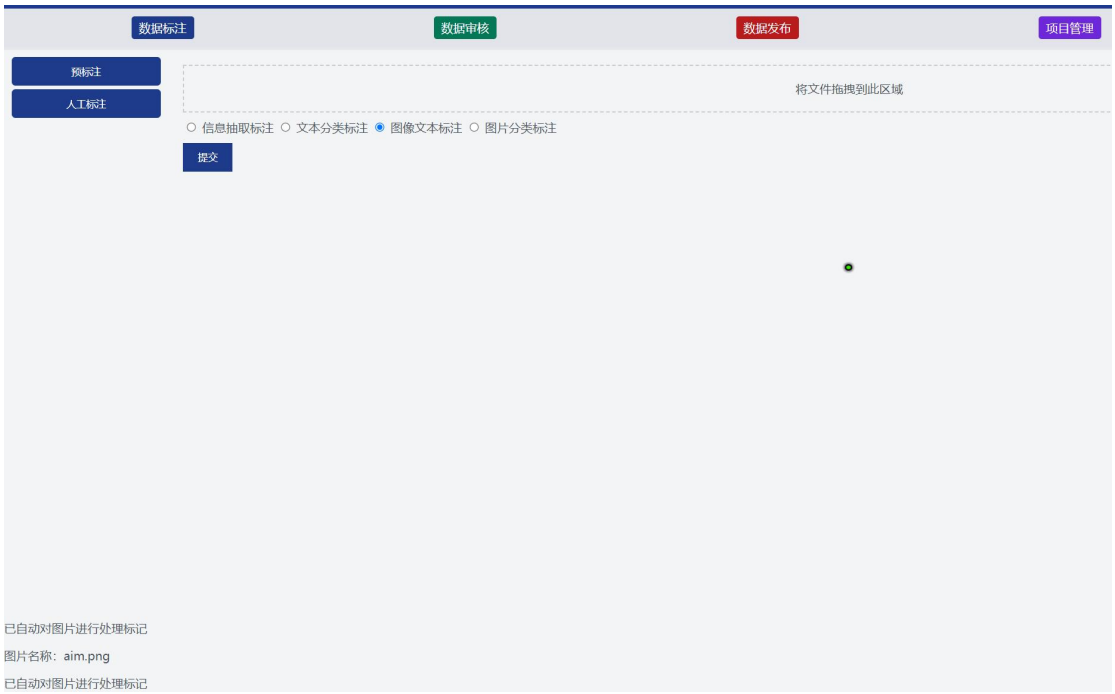

图 3- 3 图像文本标注

图 3- 4 图片分类标注

4. 后端设计

4.1. Tomcat 服务器

Tomcat 是一个成熟的、广泛使用的开源 Java Servlet 容器，已经在生产环境中被广泛使用多年。这意味着它经过了大规模的测试和验证，具有稳定性和可靠性。

此外还有一些选择它的理由：

- Java EE 支持：Tomcat 支持 Java EE 规范（如 Servlet、JSP），可以方便地托管基于这些技术的应用程序。
- 成熟的生态系统： Tomcat 有一个庞大的生态系统和活跃的社区支持，可以轻松找到大量的插件、扩展和文档资源，这些资源有助于更好地构建和维护平台。
- 复杂性和配置： Tomcat 提供了广泛的配置选项，允许管理员根据具体需求进行定制。
- 稳定性和可靠性： Tomcat 在长时间运行和生产环境中表现出了出色的稳定性和可靠性。这对于确保数据标注平台的稳定运行至关重要。

5. 数据库设计

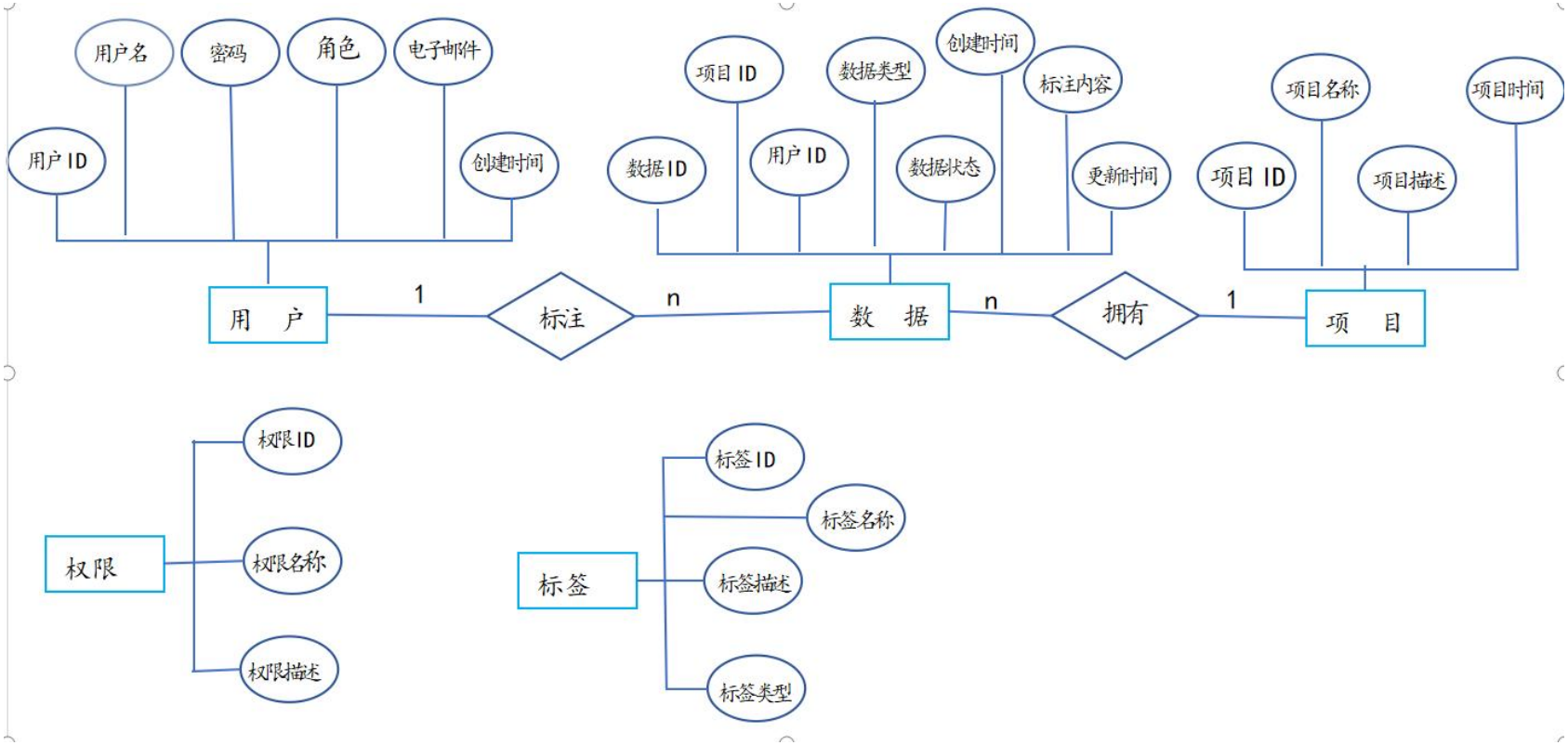

表 5- 1 E-R 图

表 5- 2 数据库设计总表

数据库设计总表			
用户表（user)			
字段	类型	主键	说明
user_id	Int（自增长）	√	唯一标识用户的 ID
username	varchar(20)		用户的用户名
password	varchar(40)		用户的密码
role	枚举		用户角色
email	varchar(50)		用户的电子邮件地址
create_time	日期时间		用户创建时间
数据标注表（data_annotation)			
字段	类型	主键	说明
data_id	Int（自增长）	√	唯一标识标注数据的 ID
project_id	外键		关联到项目的 ID
user_id	外键		关联到用户的 ID
data_type	枚举		标注数据的类型
status	枚举		标注数据的状态
content	text		存储标注内容
create_time	日期时间		标注数据创建时间
update_time	日期时间		标注数据最后更新时间
项目表（project):			
字段	类型	主键	说明
project_id	Int（自增长）	√	唯一标识项目的 ID
project_name	varchar(40)		项目名称
description	text		项目描述
create_time	日期时间		项目创建时间
权限表（permission):			
permission_id	Int（自增长）	√	唯一标识权限的 ID
permission_name	varchar(40)		权限的名称
description	text		权限的描述
标签表（label):			
label_id	Int（自增长）	√	唯一标识标签的 ID
label_name	文本		标签的名称
label_description	文本		标签的描述
data_type	枚举		标签所属的数据类型

6. 服务器设计

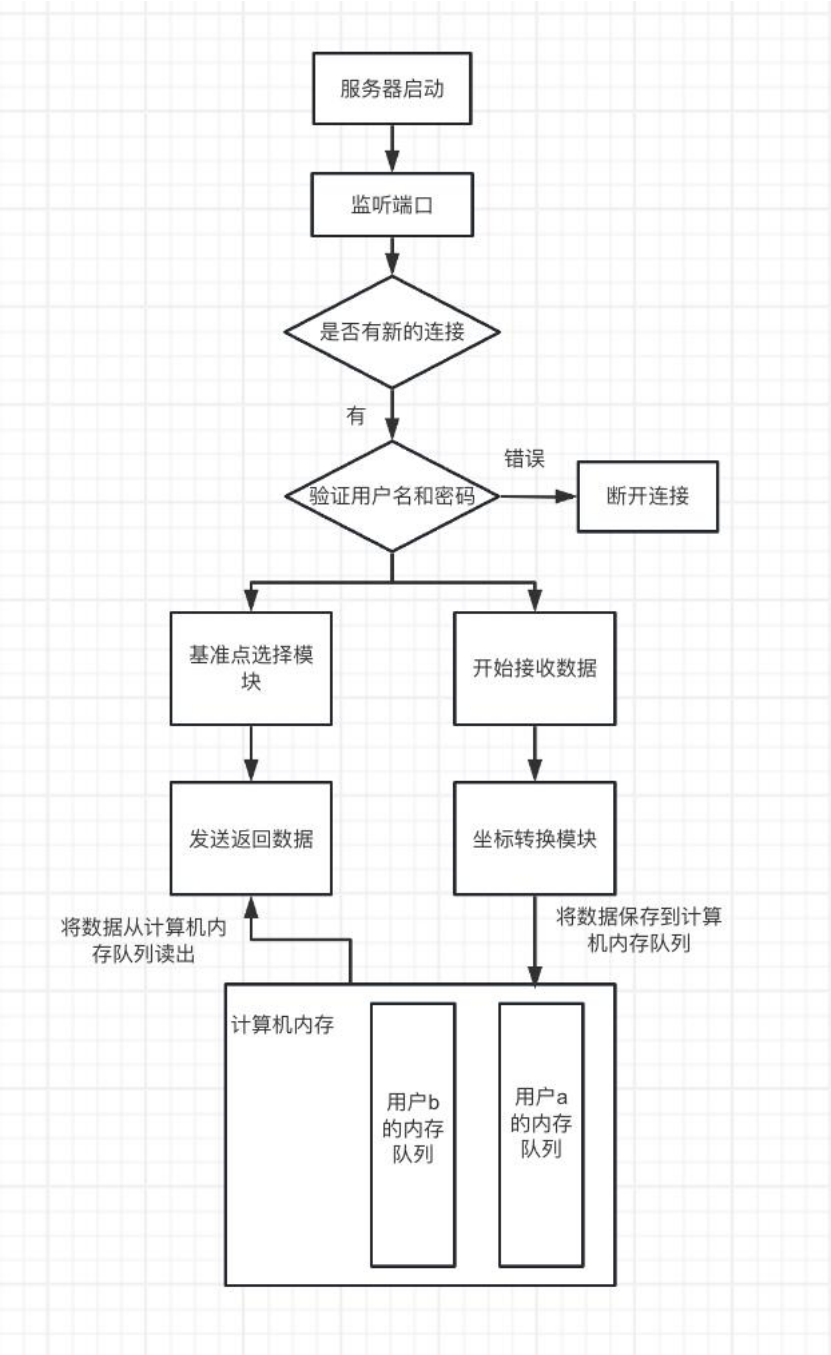

图 6- 1 服务器架构图