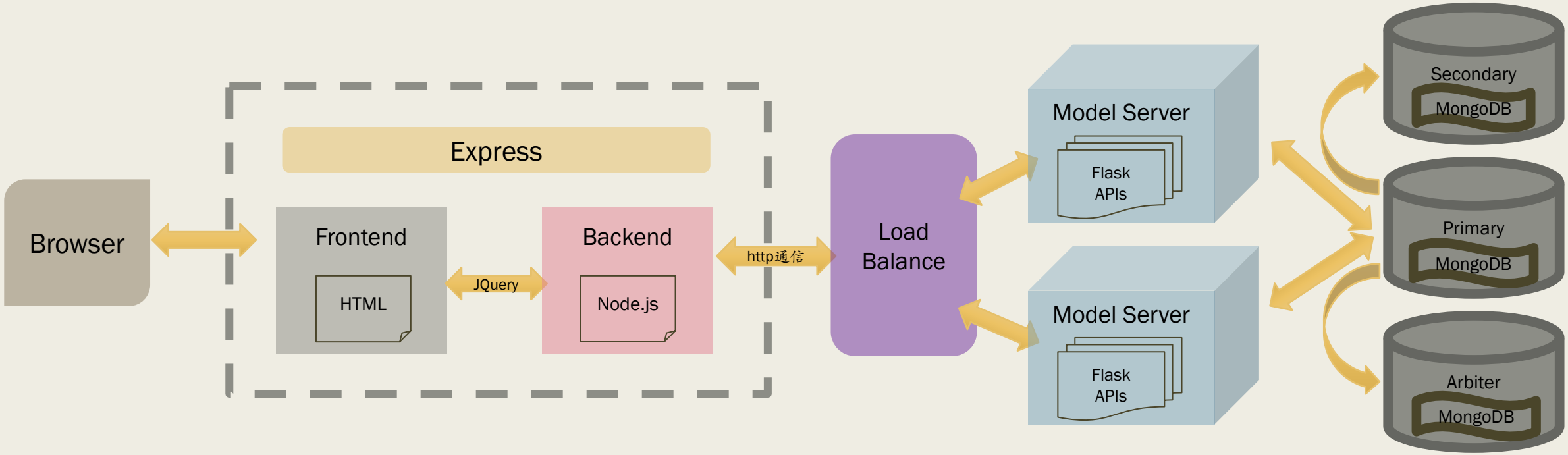
PDF TABLE DETECTION

印张悦 牛悦安 梅佳奕

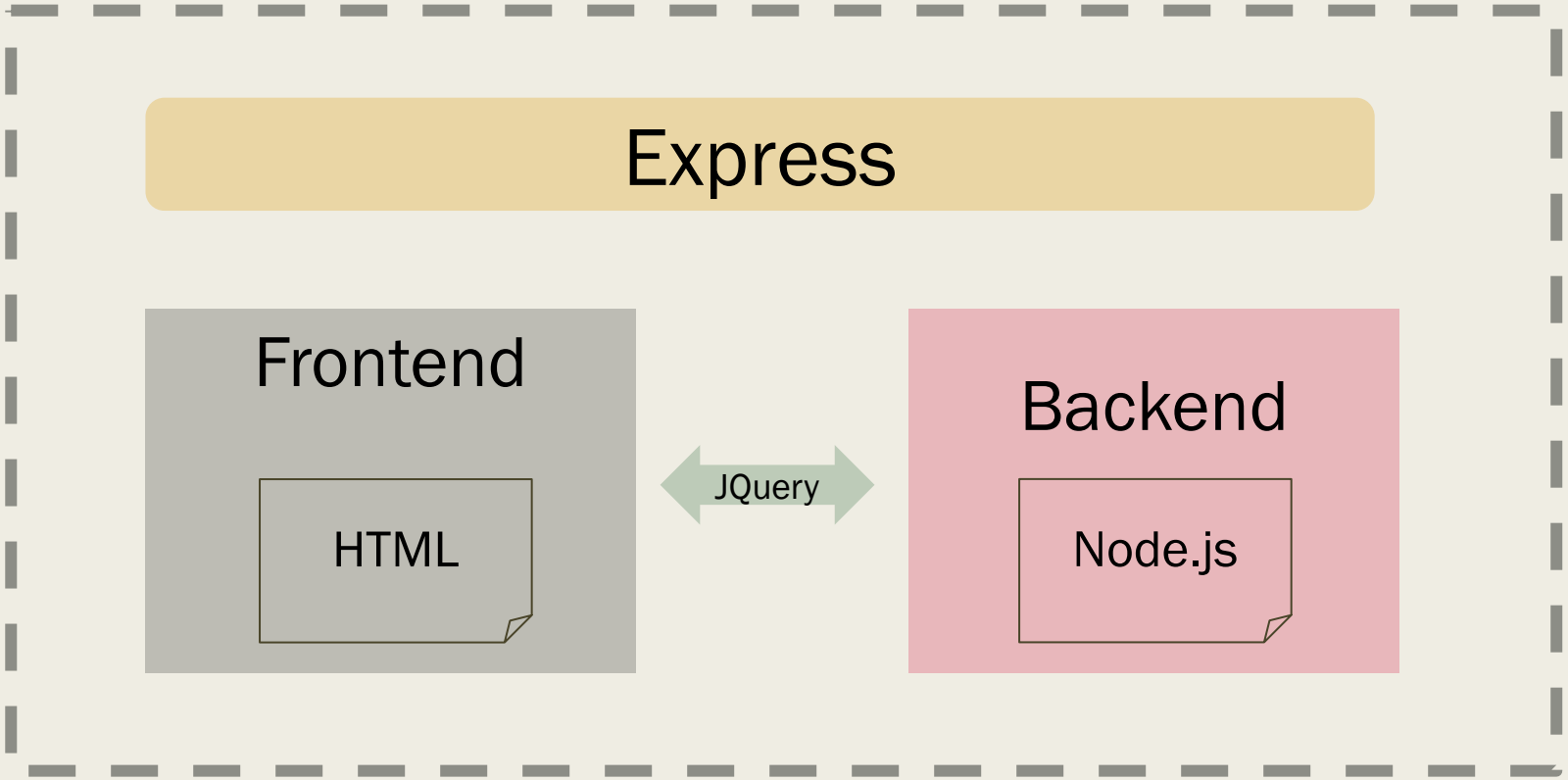
项目简介

- PDF表格提取

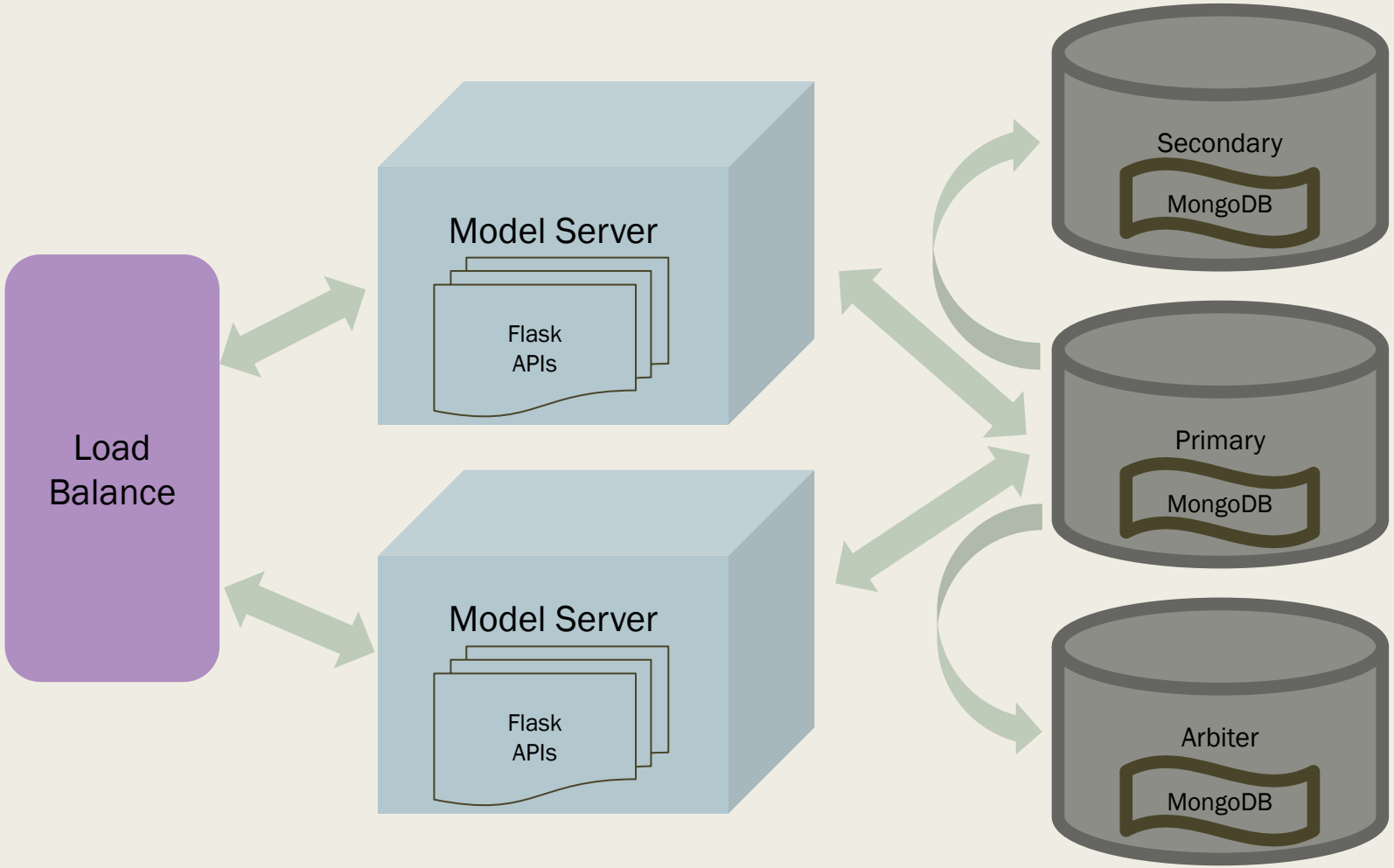
项目框架


Browser
Express
Frontend
HTML
JQuery
Backend
Node.js
http通信
Load Balance
Model Server
Flask APIs
Model Server
Flask APIs
Secondary
MongoDB
Primary
MongoDB
Arbiter
MongoDB

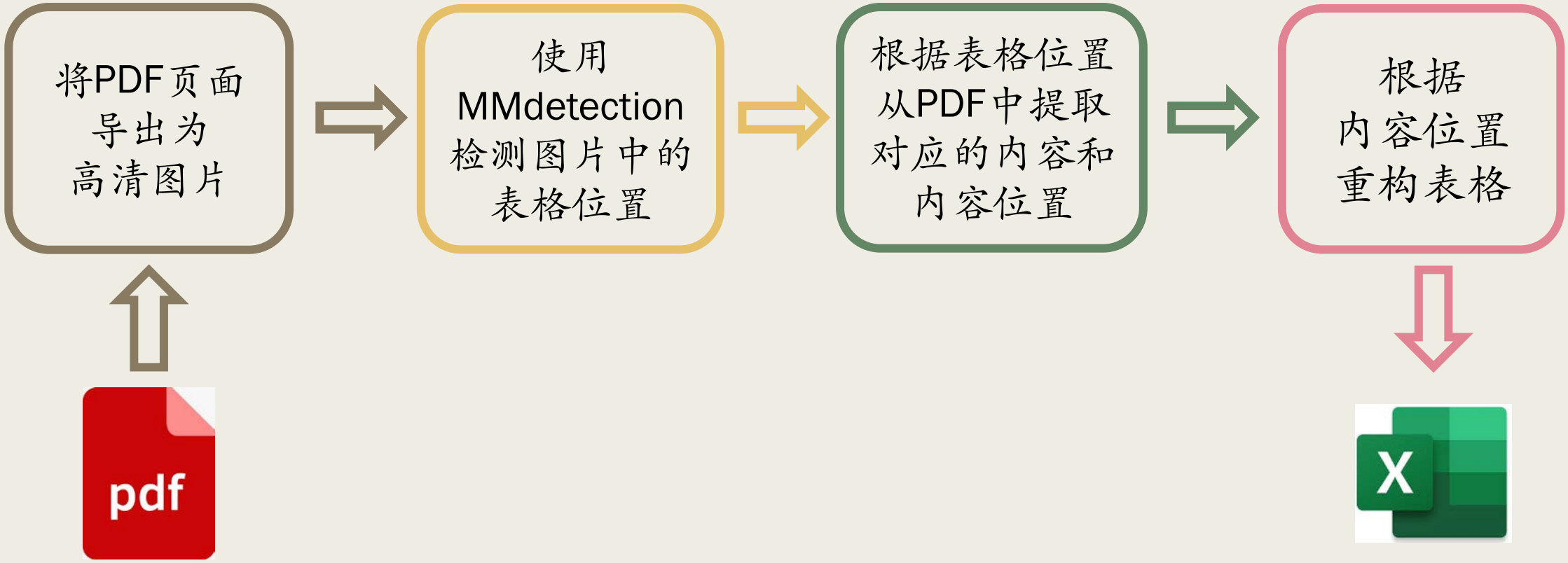
前端实现细节



后端实现细节


Load
Balance
Model Server
Flask
APIs
Model Server
Flask
APIs
Secondary
MongoDB
Primary
MongoDB
Arbiter
MongoDB

模型结构


将PDF页面
导出为
高清图片
使用
MMdetection
检测图片中的
表格位置
根据表格位置
从PDF中提取
对应的内容和
内容位置
根据
内容位置
重构表格
pdf

模型介绍

- 模型训练
  - 基于目标识别模型MMdetection，在数据集ICDAR2013、ICDAR2019进行训练，提取表格位置，单元格位置
  - 利用PyMuPDF解析pdf页面中的内容
  - 利用单元格位置以及pdf的元信息还原表格，生成xlsx

- 模型推理
  - 将MMdetection迁移到CPU进行推理

云上实际部署情况

- 前端服务器配置
  - *云主机 *1*
- 后端服务器配置
  - *负载均衡 *1*
  - *模型：云主机 *2*
- 数据库服务器配置
  - *MongoDB *3*

Demo演示

- 网址：http://106.75.237.104:3000/index.html
- 水杉在线仓库链接：http://gitea.shuishan.net.cn/10174503110/pdfTableDetection
- 演示视频链接：https://pan.baidu.com/s/1H4TsqELVMQgOzRXX-qCCWw
  - *提取码：1234*

未来展望

- 功能方面
  - *用户注册、登录*
  - *查看历史记录*
- 模型方面：
  - *模型蒸馏，压缩模型大小、加快推理速度*