数据采集子系统——项目设计报告

概述

该子系统的数据来源是国家一级博物馆的官网。爬取内容包括博物馆基本的介绍、参观信息（开放时间等）、展览信息、教育活动、经典藏品信息、学术研究信息等。对爬取信息进行加工，并存入数据库，将爬虫部署到服务器上并更新数据库。

系统设计前提

- 对系统进行全面的定位，具体实现功能的陈列
- 了解相关技术的实现，指明学习开发方向

系统设计目的

- 为其他开发组提供数据支持
- 作为提供数据支持组，要在其他组开发前完成数据上传
- 与其他组对接协调

开发工具

- Pycharm（爬虫编写工具）
- 浏览器（博物馆网站浏览）
- Navicat（数据库管理）
- Xshell（服务器管理）

开发技术

- Scrapy框架、Python、sql语句

系统运行环境

- 系统运行硬件环境
  - 普通PC机：Windows10系统
  - 服务器：CentOS系统（CPU:2.0GHz,内存:2GB及以上）
- 系统运行软件环境
  - 操作系统：Win10
  - 软件虚拟环境：Python 3
  - 数据库：MySQL
  - 浏览器：Google Chrome、Firefox、Microsoft Edge

后端服务器地址

远程连接数据库管理系统：MySQL
Ip地址：123.56.13.242
数据库端口：3306
username：root
password：Aliyun2021
数据库：museum
表：Collection、Exhibition、Museum

代码实现：

```
host="123.56.13.242",
port=3306,
user="root",
passwd="Aliyun2021",
db="museum",
charset="utf8"
```

数据模型

1、博物馆基本信息表

字段名	博物馆编号	博物馆名称	博物馆图片	博物馆评分	博物馆开放时间	博物馆地址	博物馆网址	博物馆电话	博物
变量名	mus_id	mus_name	mus_picture	mus_grade	mus_time	mus_address	mus_remark	mus_phone	mus_
数据类型	int	varchar	varchar	double	varchar	text	text	char	var

2、博物馆藏品表

字段名	藏品编号	博物馆编号	藏品名称	年代	基本介绍	博物馆名称	藏品图片
变量名	col_id	mus_id	col_name	col_era	col_info	mus_name	col_picture
数据类型	int	int	varchar	varchar	text	varchar	varchar

3、博物馆展览表

字段名	展览编号	博物馆编号	展览名称	展览内容	博物馆名称	展览图片	展览时间
变量名	col_id	mus_id	col_name	col_era	col_info	mus_name	col_picture
数据类型	int	int	varchar	text	varchar	varchar	varchar