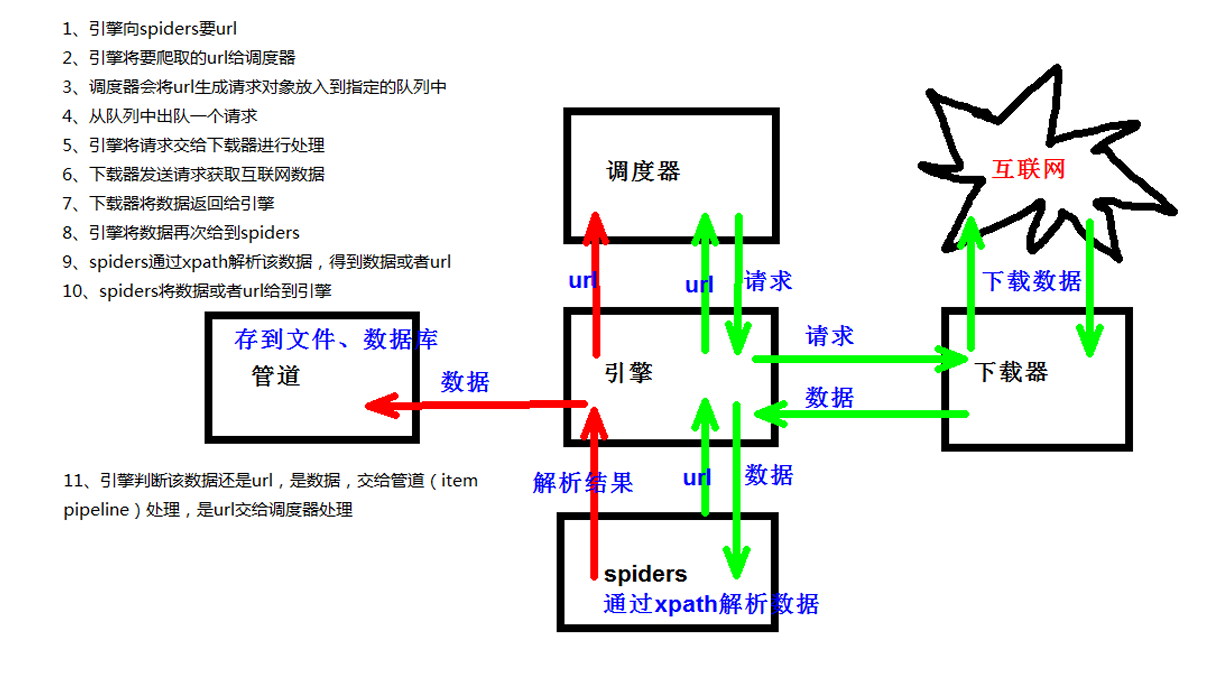
**1. 概述**

* **目的**：爬取读书网小说分栏的所有书籍信息。
* **功能**：爬虫的主要功能：读取每本数据的题目title、作者author以及展示图片src；将读取数据保存为json格式的字符串；将数据同时利用pysql进行入库。

**2. 架构设计**

* **系统架构**：

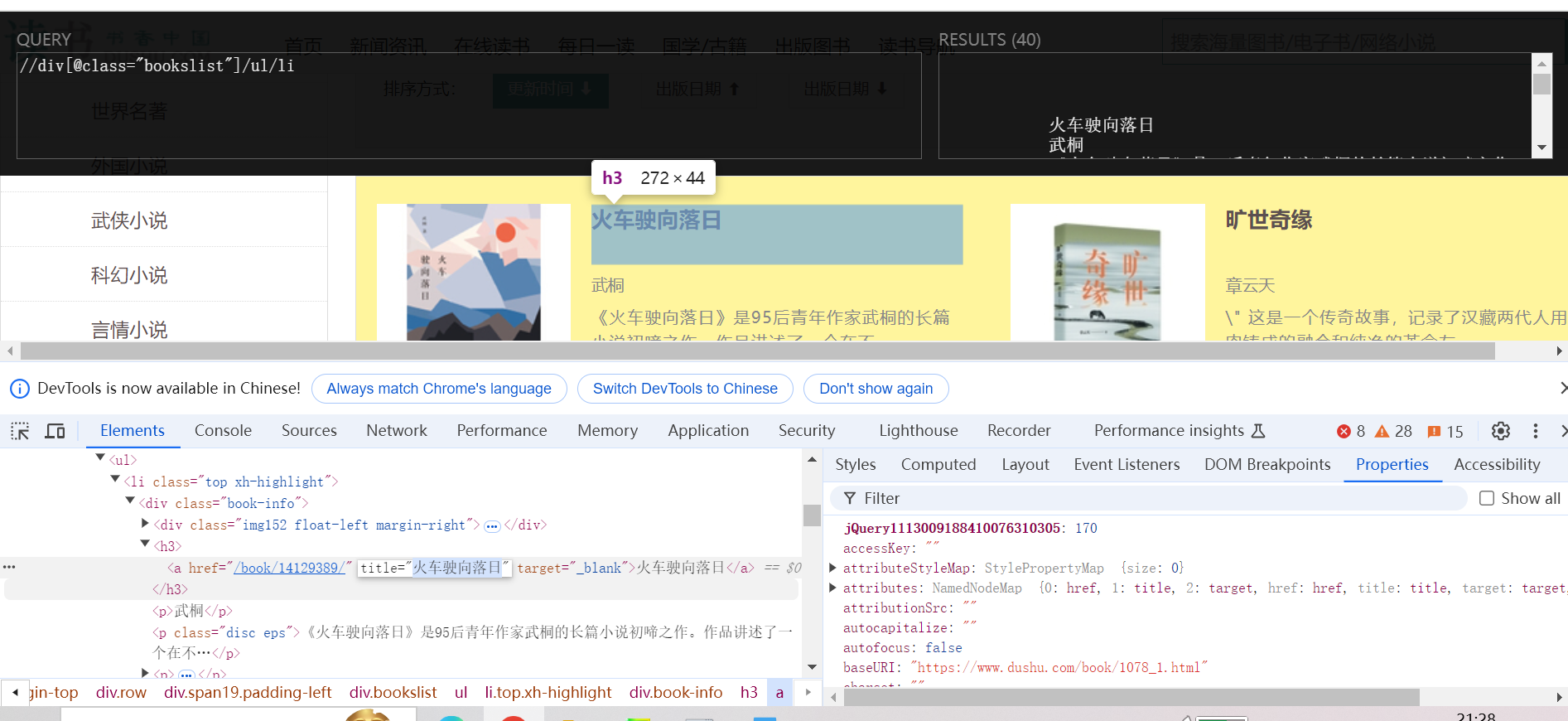


* **组件描述**：
  + **爬虫模块**：包括 spiders、middlewares、pipelines 等
  + **数据存储**：存储爬取到的数据（MySQL DB）
  + **配置管理**：配置文件 (settings.py)

**3. 数据流程**

* **数据获取**：ReadSpider 爬取书籍列表页面，并提取书籍信息。
* **数据处理**：在 parse\_item 方法中利用xpath解析 HTML 页面，提取数据。

利用xpath helper浏览器组件辅助解析数据



* **数据存储**：通过 MysqlPipeline 将数据存储到 MySQL 数据库中的 books 表。

4. **数据库设计**

CREATE TABLE books (

title VARCHAR(255),

author VARCHAR(255),

src VARCHAR(255)

);

**关系**：单表存储书籍信息，无其他表格关系。

5. **配置**

# Database settings

DB\_HOST = 'localhost'

DB\_PORT = 3306

DB\_USER = 'root'

DB\_PASSWORD = 'leo'

DB\_NAME = 'spider'

DB\_CHARSET = 'utf8'