

<基于 Hive 的数据爬取分析取系统>

概要设计

学号 姓名：

181110315 王少博

181110110 刘凯华

181110422 邹铭烨

181110406 李开新

181110316 王钟杰

目录

[一、文档介绍 1](#_Toc47594003)

[1、目的 1](#_Toc47594004)

[二、系统架构图 2](#_Toc47594005)

[1、应用结构 2](#_Toc47594006)

[2、部署方案 3](#_Toc47594007)

[3、系统交互 3](#_Toc47594008)

[三、软件包构成 3](#_Toc47594009)

[1、开发环境软件包构成 3](#_Toc47594010)

[2、测试环境软件包构成 4](#_Toc47594011)

[3、生产环境软件包构成 5](#_Toc47594012)

[四、技术架构图 6](#_Toc47594013)

[五、项目工程构成 6](#_Toc47594014)

[1、模块构成 6](#_Toc47594015)

[2、模块的文件夹构成 7](#_Toc47594016)

[3、日志文件构成 10](#_Toc47594017)

[4、Release目录构成 11](#_Toc47594018)

[5、Package构成 11](#_Toc47594019)

[5.1、Package定义规则 11](#_Toc47594020)

[5.2、Package定义示例 11](#_Toc47594021)

[六、系统通信构成 11](#_Toc47594022)

[七、数据库设计 12](#_Toc47594023)

[八、功能性架构管理 12](#_Toc47594024)

[1、日志输出规范 12](#_Toc47594025)

[1.1、业务日志输出规范 12](#_Toc47594026)

[1.2、异常日志输出规范 12](#_Toc47594027)

[2、异常处理规范 12](#_Toc47594028)

[3、数据库连接方式 13](#_Toc47594029)

[4、事务管理方式 13](#_Toc47594030)

[5、国际化 13](#_Toc47594031)

[6、归档 13](#_Toc47594032)

[7、并发管理方式 13](#_Toc47594033)

[8、负载均衡方案 13](#_Toc47594034)

[9、容灾方案 13](#_Toc47594035)

[八、非功能性架构管理 14](#_Toc47594036)

[1、网络性能保障方案 14](#_Toc47594037)

[2、安全性保障方案 14](#_Toc47594038)

[3、可维护性保障方式 14](#_Toc47594039)

[4、可扩展性保障方式 14](#_Toc47594040)

[九、服务监控方案 14](#_Toc47594041)

# 文档介绍

## 1、目的

本文档记录股票爬虫系统的概要设计信息。主要包括对系统的概述，对系统体系结构、系统共有的功能，以及系统配置的设计。该文档作为系统的设计人员以及开发人员设计开发的主要参考依据。

2、 术语：

Hadoop：Linux集群环境

Hive：基于Hadoop的数据仓库平台

Django：python web开发框架

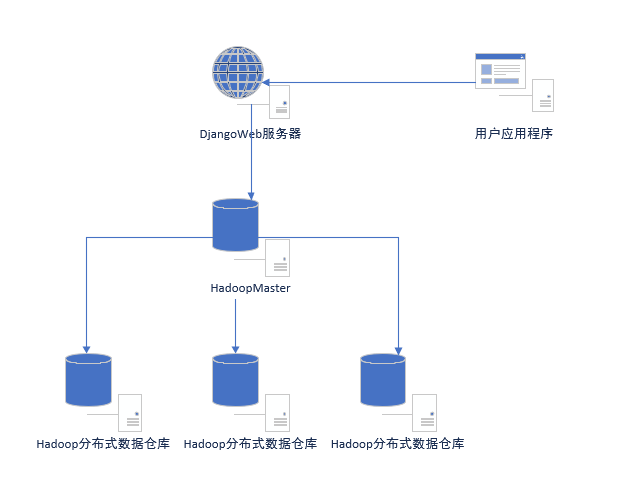
PyEcharts：开源第三方库

# 系统架构图

## 1、应用结构



## 2、部署方案



## 3、系统交互

组件（子系统）逻辑图

# 软件包构成

## 1、开发环境软件包构成

|  |  |  |
| --- | --- | --- |
| 软件包 | 版本 | 所有者 |
| Mysql | 5.7.19-1 | Oracle |
| Django | 2.7 | 开源 |
| Hadoop | 2.7.2 | ApacheSoftware Foundation |
| Java | 1.8.1 | Oracle |
| Python | 3.7 | 开源 |
| PyEcharts | V1.0 | 开源 |
| Hive | 1.2.1 | ApacheSoftware Foundation |

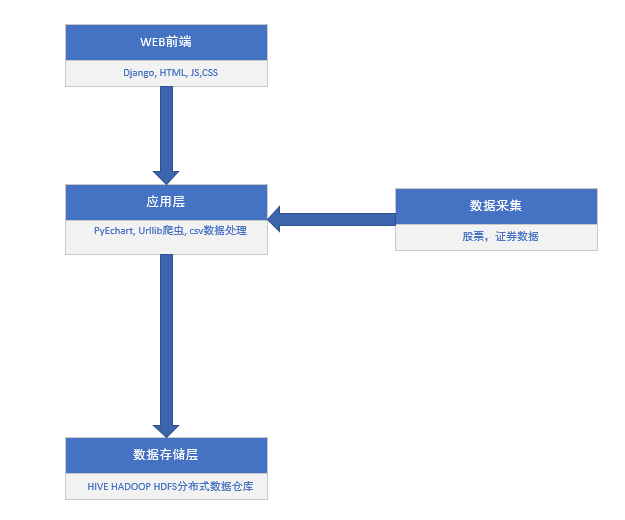
## 2、测试环境软件包构成

|  |  |  |
| --- | --- | --- |
| 软件包 | 版本 | 所有者 |
| Mysql | 5.7.19-1 | Oracle |
| Django | 2.7 | 开源 |
| Hadoop | 2.7.2 | ApacheSoftware Foundation |
| Java | 1.8.1 | Oracle |
| Python | 3.7 | 开源 |
| PyEcharts | V1.0 | 开源 |
| Hive | 1.2.1 | ApacheSoftware Foundation |

## 3、生产环境软件包构成

|  |  |  |
| --- | --- | --- |
| 软件包 | 版本 | 所有者 |
| Mysql | 5.7.19-1 | Oracle |
| Django | 2.7 | 开源 |
| Hadoop | 2.7.2 | ApacheSoftware Foundation |
| Java | 1.8.1 | Oracle |
| Python | 3.7 | 开源 |
| PyEcharts | V1.0 | 开源 |
| Hive | 1.2.1 | ApacheSoftware Foundation |

# 技术架构图



# 项目工程构成

## 1、模块构成

后端：

Back-server

├─server-api

├─server -common

└─server -web

前端：

Front-UI

├─css

├─fonts

├─images

├─js

├─libs

├─mobile

├─modules

└─plugins

## 2、模块的文件夹构成

前端：

Financialdatadisplay

├─manage.py

├─db.sqlite3

├─financialdatadisplay

├─├─\_init\_.py

├─├─settings.py

├─├─urls.py

├─├─wsgi.py

├─fdd

├─├─\_pycache\_

├─├─miggrations

├─├─\_init\_.py

├─├─admin.py

├─├─app.py

├─├─models.py

├─├─tests.py

├─├─urls.py

├─├─views.py

├─├─templates

├─├─├─fdd

├─├─├─├─home.html

├─├─├─├─navbar.html

├─├─├─├─aboutus.html

├─├─├─├─白银.html

├─├─├─├─晨光.html

├─├─├─├─大豆.html

├─├─├─├─海康.html

├─├─├─├─黄金.html

├─├─├─├─金山.html

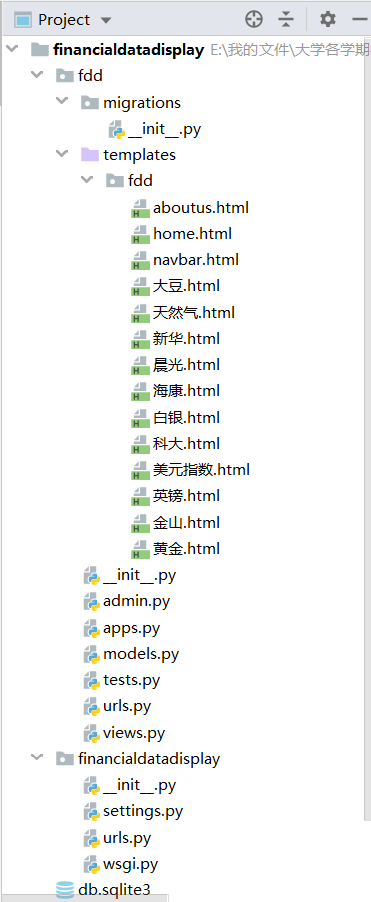
├─├─├─├─科大.html

├─├─├─├─美元指数.html

├─├─├─├─天然气.html

├─├─├─├─新华.html

├─├─├─├─英镑.html



## 3、日志文件构成

系统日志放在/logs/下面，按天进行压缩后转存。

其中也保存着hadoop的namenode和datanode相关日志，可用来协助处理开发过程中的意外报错

## 4、Release目录构成

<各模块在生产环境中部署的位置、权限（如linux上的755）、所有者（如root），所属组（如root）、用途等>

## 5、Package构成

### 5.1、Package定义规则

1~2层是domain名称为：com.your-company-name，3层按框架模块命名，4层按功能模块命名，5层按MVC分层命名

### 5.2、Package定义示例

cn.com.dhc.business.deviceinfo.controller/DeviceInfoController.java

# 系统通信构成

服务器通讯是通过HTTP接口实现，发送HTTP请求和HTTP响应。

与数据仓库进行通讯使用HIVE，HADOOP的通讯接口。在WINDOWS上使用impala，thrift与数据仓库进行连接。

爬取网页信息采用Urllib/requests库，通过发送HTTP请求获取信息。

除了使用数据仓库进行存储外，中间件也使用了数据库，csv文件加快存取速度。

# 数据库设计

# 八、功能性架构管理

## 1、日志输出规范

引用《PD-IT-2017522-开发规范.04.日志规范.v1.0.0》

### 1.1、业务日志输出规范

在controller中的方法进入及离开都要加上日志记录，如果参数中有业务编码则将业务编码输出到日志中。

代码中如果有中断，则在中断之前也要有日志输出。

### 1.2、异常日志输出规范

每个方法都做统一异常拦截，统一异常中输出异常日志。

Try保护的异常在catch中将异常打印到日志。

## 2、异常处理规范

发生异常后，要做日志记录、事物回滚(如有)及用户界面的友好提示。

## 3、数据库连接方式

采取Hadoop，Hive的HDFS分布式数据仓库进行数据存储。

使用impala.dbapi与Hive进行连接。

使用pymssql/pymysql/sqlite3与数据库MySql，SqlServer，Sqlite进行连接。

## 4、事务管理方式

采取Hive自带的事务管理方式。

## 5、国际化

和国际接轨，爬取股票指数。

## 6、归档

无

## 7、并发管理方式

采取合作开发，不设定小组领导，小组成员共同讨论探讨进行开发。

## 8、负载均衡方案

使用Hadoop集群进行负载平衡

## 9、容灾方案

无。

# 非功能性架构管理

## 网络性能保障方案

* 1. 采取内部局域网互联

## 安全性保障方案

* 1. 设置安全密钥

## 可维护性保障方式

* 1. 记录错误信息，用于维护

## 可扩展性保障方式

* 1. 模块之间相互独立，用于增强可扩展性。

# 服务监控方案

无