# 摘 要

随着计算机技术的迅猛发展，许多行业、单位或机构都逐步实现了业务、信息的计算机化管理。在企业实施信息化进程中，由于企业内各部门的业务和功能归属不同，以及各企业信息化水平存在差异等原因，造成企业拥有许多相互隔离的、由不同核心技术构建的信息服务和管理系统,从而使得企业内以及企业间的协作效率与效果很不理想。异构数据以其复杂性、动态变化和不确定性给系统间的数据共享造成极大障碍，导致信息孤岛现象日益严重。由于各部门信息化程度及系统封闭性的原因，集成系统面临许多技术障碍。而数据集成技术可以将多个分散的、异构的数据源统一起来，为用户提供统一的数据访问接口，实现异构数据的共享。

本系统分为数据解析处理和数据存储两大部分，数据解析处理部分由JAVASE来完成，其中用到了字符串解析和XML解析。数据存储部分选择将数据解析处理部分结构化的数据文件load进Hive进行存储。原始数据选定了微博的用户基本信息和用户评论信息，用户基本信息数据格式为JSON，用户评论信息数据格式为XML。

**关键词**：异构数据；数据的共享利用；信息孤岛；XML解析；JAVASE；Hive

ABSTRACT

With the rapid development of computer technology, many industries and organizations or institutions gradually realized the computerization of business and information management. In the process of enterprise informatization, because of the different departments within the enterprise business and belonging to different functions, and the reason of the difference of each enterprise informatization level, enterprises have caused many isolated, built by different core technology of information service and management system, so as to make the enterprise and enterprise collaboration between the efficiency and the effect is not ideal. Heterogeneous data for its complexity, dynamic changes and uncertainty to the system data sharing between the huge obstacles, lead to serious phenomenon of information island. Due to the degree of information and system closure of each department, the integrated system is faced with many technical obstacles. Data integration technology can unify multiple scattered and heterogeneous data sources, provide unified data access interface for users, and realize heterogeneous data sharing.

The system is divided into two parts: data parsing and data storage. The part of data parsing is done by JAVASE, which USES string parsing and XML parsing. The data storage section chooses to load the structured data files of the data parsing processing section into hive for storage. The original data selects the basic user information and user comment information of WEIBO. The format of the basic user information data is JSON, and the format of the user comment information data is XML.

**Key words:** Heterogeneous data; Shared utilization of data; Information islands;

XML parsing. JAVASE; Hive

目 录

第一章 *绪论*……………………………………………………………1

1.1背景………………………………………………………………………………1

1.2相关领域国内外发展…………………………………………………………×

1.3研究内容………………………………………………………**………………**×

1.4论文的组织结构………………………………………………**………………**×

第二章 系统技术综述…………………………………………………×

2.1数据解析部分技术**……………………………………………………………**×

2.1.1 JSON类型数据解析技术**……………………………………………**×

2.1.2 XML类型数据解析技术**………………………………………………**×

2.2数据存储部分技术**……………………………………………………………**×

2.2.1 HADOOP**…………………………………………………………………**×

2.2.2 HIVE**……………………………………………………………………**×

2.3开发工具介绍**…………………………………………………………………**×

2.3.1 Eclipse**………………………………………………………………**×

2.3.2 VMware Workstation**…………………………………………………**×

第三章 系统需求与可行性分析………………………………………×

3.1用户需求分析**…………………………………………………………………**×

3.2系统需求分析**…………………………………………………………………**×

3.3性能需求分析**…………………………………………………………………**×

3.4系统可行性分析**………………………………………………………………**×

第四章 异构数据集成系统设计………………………………………×

4.1☆☆☆☆**………………………………………………………………………**×

4.2☆☆☆☆**………………………………………………………………………**×

第五章 异构数据集成系统实现………………………………………×

5.1☆☆☆☆**………………………………………………………………………**×

5.2☆☆☆☆**………………………………………………………………………**×

第六章 测试……………………………………………………………×

5.1☆☆☆☆**………………………………………………………………………**×

5.2☆☆☆☆**………………………………………………………………………**×

结束语……………………………………………………………………×

致谢………………………………………………………………………×

参考文献…………………………………………………………………×

附录……………………………………………………………………×

*一级目录1.5倍行距，段前0.5倍行距*

*二、三级目录单倍行距，段前段后为0*

# 第一章 绪论

## 1.1背景

随着计算机技术的迅猛发展，许多行业、单位或机构都逐步实现了业务、信息的计算机化管理。在企业实施信息化进程中，由于企业内各部门的业务和功能归属不同，以及各企业信息化水平存在差异等原因，造成企业拥有许多相互隔离的、由不同核心技术构建的信息服务和管理系统,从而使得企业内以及企业间的协作效率与效果很不理想。异构数据以其复杂性、动态变化和不确定性给系统间的数据共享造成极大障碍，导致信息孤岛现象日益严重。由于各部门信息化程度及系统封闭性的原因，集成系统面临许多技术障碍。而数据集成技术可以将多个分散的、异构的数据源统一起来，为用户提供统一的数据访问接口，实现异构数据的共享。

# 第二章 ☆☆☆☆☆

*（章号与章名间半角空格两个，居中、三号、黑体）*

## 2.1☆☆☆（*四号、黑体、顶格*）

### 2.1.1☆☆☆*（四号、黑体、顶格）*

□□☆☆☆☆☆☆☆☆☆正文*（小四号、宋体、空2个字符、用1.25倍行间距）*

……….

……..

……..

……..

……..

……..

……..

……..

……..

图2.5 **×××**结构图

*半角空格两个，宋体五号居中，位于图下*

*图与下文空一行，图名必须与图在一页，图不允许分页排版，图中字体五号，汉字宋体，英文Times New Roman*

图像的灰度处理总体分为以下三种方法来实现：

第一种是首先获取到该图像每个像素的R、G、B分量，计算它们的平均值之后，将它赋予每个分量。用数学表达式为：

Gray=（R+G+B）/3 公式（2.1）

*公式按章依次编*

*号，中间用.分割*

*公式前后与正文间隔6磅*

2

# 结束语

*（居中、三号、黑体，段前、后1行，1.25倍行距）*

（正文内容：小四号宋体，行距1.25倍）

*论述项目完成情况，优缺点，实用性，展望等与项目有关的内容，并非心得。*

3

# 致 谢

*（居中、三号、黑体，段前、后1行，1.25倍行距、致谢两字空半角空格两格）*

（正文内容：小四号宋体，行距1.25倍，200字以内）

4

# 参考文献*（三号、黑体、顶格，段前、后1行，1.25倍行距）*

[1] [凤祥云](http://xueshu.baidu.com/s?wd=author%3A%28%E5%87%A4%E7%A5%A5%E4%BA%91%29%20&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight%3Dperson" \t "_blank)，[孙海艳](http://xueshu.baidu.com/s?wd=author%3A%28%E5%AD%99%E6%B5%B7%E8%89%B3%29%20&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight%3Dperson" \t "_blank)，[张万臣](http://xueshu.baidu.com/s?wd=author%3A%28%E5%BC%A0%E4%B8%87%E8%87%A3%29%20&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight%3Dperson" \t "_blank).[基于光纤通信技术的物联网传感器系统](http://xueshu.baidu.com/s?wd=paperuri%3A%28c77a9549c24b8365681bc6195aed1b3c%29&filter=sc_long_sign&tn=SE_xueshusource_2kduw22v&sc_vurl=http%3A%2F%2Fwww.cqvip.com%2FQK%2F91041X%2F201607%2F669514878.html&ie=utf-8&sc_us=7665783302043157044" \t "_blank)[J].激光杂志, 2016(7):131-134．***期刊请写明页码***

[2] [迪内希·钱德拉·维玛](http://xueshu.baidu.com/s?wd=author%3A%28%E8%BF%AA%E5%86%85%E5%B8%8C%C2%B7%E9%92%B1%E5%BE%B7%E6%8B%89%C2%B7%E7%BB%B4%E7%8E%9B%29%20&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight%3Dperson" \t "_blank)，[帕利德·维玛](http://xueshu.baidu.com/s?wd=author%3A%28%E5%B8%95%E5%88%A9%E5%BE%B7%C2%B7%E7%BB%B4%E7%8E%9B%29%20&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight%3Dperson" \t "_blank).大数据爆炸时代的移动通信技术与应用[M]. 郎为民,译．北京：机械工业出版社, 2016：20-30．***书籍请写明出版社、页码***

[3] [刘俊文](http://xueshu.baidu.com/s?wd=author%3A%28%E5%88%98%E4%BF%8A%E6%96%87%29%20&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight%3Dperson" \t "_blank)，[赵子岩](http://xueshu.baidu.com/s?wd=author%3A%28%E8%B5%B5%E5%AD%90%E5%B2%A9%29%20&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight%3Dperson" \t "_blank)，[徐慧明](http://xueshu.baidu.com/s?wd=author%3A%28%E5%BE%90%E6%85%A7%E6%98%8E%29%20&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight%3Dperson" \t "_blank)，[张素香](http://xueshu.baidu.com/s?wd=author%3A%28%E5%BC%A0%E7%B4%A0%E9%A6%99%29%20&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight%3Dperson" \t "_blank)．[量子通信技术在电力信息系统保密传输中的应用](http://xueshu.baidu.com/s?wd=paperuri%3A%2807b58d40f1334061b898700237944390%29&filter=sc_long_sign&tn=SE_xueshusource_2kduw22v&sc_vurl=http%3A%2F%2Fcpfd.cnki.com.cn%2FArticle%2FCPFDTOTAL-YDDX201609001069.htm&ie=utf-8&sc_us=2010724533587303457" \t "_blank)[C] [电力行业信息化年会](http://xueshu.baidu.com/usercenter/data/journal?cmd=jump&wd=confuri%3A%285b41c4bb46127d22%29%20%E7%94%B5%E5%8A%9B%E8%A1%8C%E4%B8%9A%E4%BF%A1%E6%81%AF%E5%8C%96%E5%B9%B4%E4%BC%9A&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight%3Dpublish&sort=sc_cited" \t "_blank" \o "电力行业信息化年会), 2016

[4] DL/T5344-2006,电力光纤通信工程验收规范[S].2006.

[5] 姜锡洲.一种温热外敷药制备方案：中国,881056073[P].1989-07-26.

[6] 王明亮.关于中国学术期刊标准化数据库系统工程的进展[EB/OL]. (1998-08-16)[1998-10-04].http：//[www.cajcd](http://www.cajcd).edu.cn/pub/wml.tex/980810-2.html.

[7] 丁文祥.数字革命与竞争国际化[N].中国青年报,2000-11-20（15）.

[8] 张志祥.间断动力系统的随机扰动及其在守恒律方程中的应用[D].北京：北京大学数理学院,1998.

[9] World Health Organization.Factors regulating the immune response:report of WHO Scientific Group[R].Geneva:WHO,1970.

*所有文献要求2010年之后，10篇以上文献，其中包含外文文献至少3篇，引用网页数量不超过文献总量的30%.所有参考文献需在正文中有引用标志）*

*中文参考文献用五号宋体，外文的用五号Times New Roman字体，均为1.25倍行距*

*一条参考文献如果占两行，第二行文字要位于序号的后边，与第一行文字对齐。*

5

5

# 附录X*(与参考文献标题格式一致，X表示A，B，C等)*

（小四号宋体，行距1.25倍）×××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××

6