拉斐尔企业画像系统

架构设计报告

|  |  |  |
| --- | --- | --- |
| 文件状态：  [√] 草稿  [ ] 正式发布  [ ] 正在修改 | 文件标识： | Company-Project-Invoice\_matchmaker |
| 当前版本： | V1.0 |
| 作 者： | 杨成凯 |
| 完成日期： | 2020-09-11 |

版 本 历 史

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 版本/状态 | 作者 | 参与者 | 起止日期 | 备注 |
| V1.0 | 杨成凯 |  | 20200911 |  |
| V1.1 | 吴月飞 |  | 20200918 |  |
| V1.2 | 杨成凯、吴月飞 | 包红雨 | 20200921 | 增加总体架构、子模块等部分。 |

目 录

[1. 需求介绍 5](#_Toc273540902)

[2. 关键问题 5](#_Toc1659588900)

[3. 关键策略 5](#_Toc1555148583)

[4. 系统总体结构 7](#_Toc1367982141)

[4.1 应用架构 7](#_Toc566718417)

[4.2 拓扑架构 8](#_Toc1195769858)

[5. 数据接口服务子模块 9](#_Toc457681052)

[5.1 主要接口 9](#_Toc1992326258)

[5.1.1精准查询接口 9](#_Toc1625632882)

[5.1.2多维查询接口 9](#_Toc34840710)

[5.1.3指标接口 9](#_Toc1341404093)

[5.1.4上下游关系图谱接口 10](#_Toc54782499)

[5.2 安全性 10](#_Toc68544378)

[5.3 其他功能 10](#_Toc2083057069)

[6. 数仓模型子模块 11](#_Toc367524872)

[6.1企业画像指标统计规则 11](#_Toc79167311)

[6.2数据存储介质 11](#_Toc95746705)

[6.2.1 企业基本信息模块 11](#_Toc1399011949)

[6.2.2 发票模块（进项发票，销项发票） 11](#_Toc480265104)

[6.2.3 交易对手模块 12](#_Toc905019358)

[6.2.3 三大会计报表模块 12](#_Toc927846551)

[6.2.4 财务模块 12](#_Toc1838932470)

[6.2.5 税务模块 12](#_Toc963686326)

[6.2.6 常用指标宽表 12](#_Toc1283324173)

[6.3关键数据脱敏 12](#_Toc342709199)

[7. 图数据库子模块 13](#_Toc2002908122)

[7.1 数据同步方式 13](#_Toc1339044172)

[7.2 图数据存储结构 13](#_Toc771371728)

[8. 开发环境的配置 13](#_Toc227753581)

[9. 运行环境的配置 13](#_Toc57821905)

[10. 测试环境的配置 14](#_Toc1444782455)

# 

# 需求介绍

见《拉斐尔企业画像报告系统\_产品需求文档》。

# 关键问题

1. 系统主要有企业的哪些数据？
2. 数据的实时性是怎样的？
3. 系统分为几个子模块？
4. 子模块主要承担了什么功能？
5. 子模块主要存储了哪些数据？
6. 子模块的数据主要经过了哪些计算？
7. 子模块主要使用了什么技术栈？
8. 子模块之间是如何进行数据交换？

9、企业的数据如何展示？

10、企业的数据如何对外提供？

11、企业的数据安全性如何保证？

12、系统的应用架构、数据架构、网络拓扑架构是怎么样的？

13、系统的计算、存储成本大约是多少？

# 关键策略

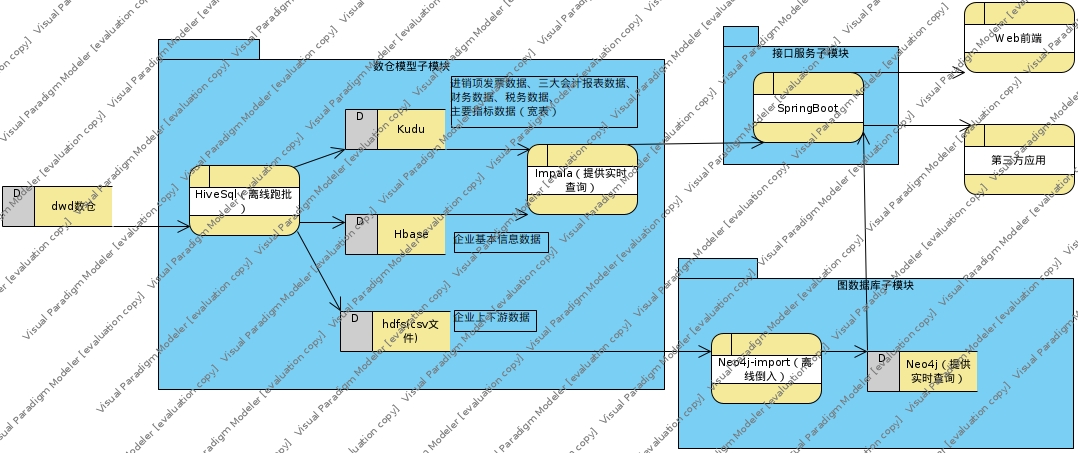
1. 系统主要有7类企业数据，企业基本信息数据、进销项发票数据、三大会计报表数据、财务数据、税务数据、主要指标数据（宽表）、企业上下游数据。详情见《企业画像指标规则\_1.0.xlsx》、《北京\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*司企业画像报告 .pdf》。
2. 以月、季、年为时间周期，计算、存储、展示上一周期的数据。
3. 系统主要有4个模块，数仓模型子模块、图数据子模块、接口服务子模块、前端展示子模块。
4. 数仓模型子模块负责生成票、财、税指标数据，并提供图数据子模块基础数据。图数据子模块主要存储关系型类图数据，并提供简单的图计算功能。接口服务子模块主要查询数仓模型子模块、图数据子模块的数据，提供外部接口。
5. 数仓模型子模块存储全部数据，图数据子模块存储企业上下游数据，接口服务子模块存储用户权限相关数据。
6. 数仓模型子模块中主要需要整合企业基本信息数据，作为“主数据”，从dw层计算票、财、税、宽表数据，根据发票数据加工企业上下游数据。图数据子模块目前不考虑历史数据与时间序列相关问题，只存储当期上下游数据用于查询。
7. 数仓模型子模块主要使用kudu、Hbase进行存储，impala提供统一查询服务，计算过程主要使用Hive sql进行。图数据子模块目前使用neo4j进行存储。接口服务子模块使用Spring Boot提供服务。
8. 数仓模型子模块与图数据子模块之间。考虑到企业的上下游数据节点数量可能会达到千万级别，图数据子模块采用每月停机维护的方式进行数据同步，数据同步以Neo4j-import的方式每月进行初始化，使用csv文本进行同步。接口服务子模块与数仓模型子模块之间，使用Spring data通过impala进行查询。接口服务子模块与图数据子模块之间，使用Spring data Neo4j直接进行查询。
9. 数据展示采用前后端分离的方式进行，后端通过接口服务子模块提供。接口服务子模块实时查询数仓模型子模块、图数据子模块。
10. 企业的查询可使用接口服务子模块，但大批量的跨部门数据同步建议使用数仓模型子模块进行数据推送。
11. 数仓模型子模块采用关键数据脱敏、服务接口子模块使用Spring Secutiy提供Api token的权限认证。
12. 见总体架构。
13. 待评估。

# 系统总体结构

系统的总体架构主要通过应用架构、拓扑架构2个方面进行考虑。

## 4.1 应用架构

本应用架构主要展示数据的存储介质、数据的流动方向。



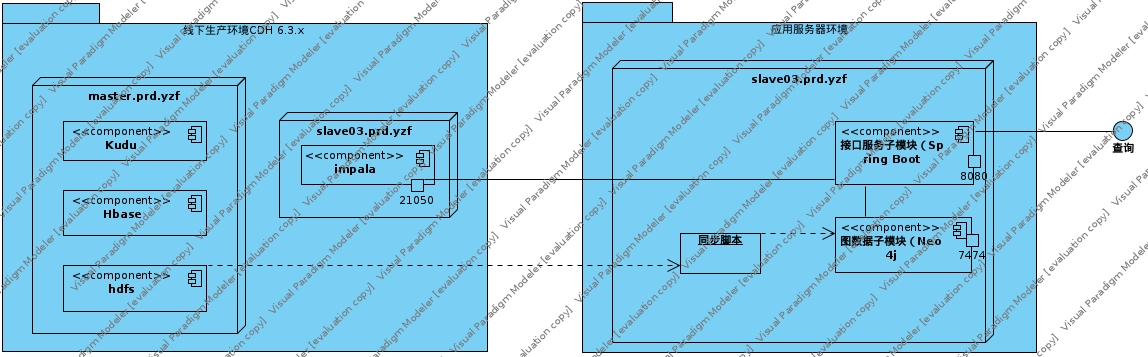
数仓模型子模块的数据来源主要来自于dwd数据仓库。数仓模型子模块按照月、季、年的周期，通过Oozie调度，将数据加工成指标。其中企业基本信息数据存储在Hbase中，进销项发票数据、三大会计报表数据、财务数据、税务数据、主要指标数据（宽表）存储在Kudu中，企业上下游数据存储在以csv格式存储在hdfs目录里中。

图数据库子模块的数据来源来自于hdfs的csv文件，数据量约在百万到千万之间。使用python脚本，通过Neo4j-import工具，将数据加载到Neo4j中。

接口服务子模块数据来源于impala，neo4j。**数仓模型子模块应处理好查询资源分配问题**，接口服务子模块只在业务逻辑上控制单一用户的查询频率。

## 4.2 拓扑架构

拓扑架构主要表示系统的运行环境。



数仓模型子模块运行在线下生产环境CDH 6.3.x中。

为不影响集群的运行，接口服务子模块、图数据子模块、同步脚本单独运行在一台应用服务器中。Neo4j以进程实例方式运行。接口服务以docker方式运行。

# 数据接口服务子模块

## 5.1 主要接口

### 5.1.1精准查询接口

根据云帐房内部企业id字段、企业全称、统一社会信用代码精确查询;根据企业简称模糊查询。

返回结果为命中的结果列表，结果包含一些企业基本信息中的关键字段。

### 5.1.2多维查询接口

至少根据代帐公司维度、企业所属地区、企业类型等维度进行查询。

返回结果为命中的结果列表，结果包含一些企业基本信息中的关键字段。

### 5.1.3指标接口

#### 5.1.3.1 企业基本信息数据接口

根据云帐房内部企业id字段，返回企业基本信息。

#### 5.1.3.2 进销项发票数据接口

根据云帐房内部企业id、帐期，返回指标数据。帐期如无指定， 默认使用最新帐期。

#### 5.1.3.3 三大会计报表数据接口

根据云帐房内部企业id、帐期，返回指标数据。帐期如无指定， 默认使用最新帐期。

#### 5.1.3.4 财务数据接口

根据云帐房内部企业id、帐期，返回指标数据。帐期如无指定， 默认使用最新帐期。

#### 5.1.3.5 税务数据接口

根据云帐房内部企业id、帐期，返回指标数据。帐期如无指定， 默认使用最新帐期。

### 5.1.4上下游关系图谱接口

根据云帐房内部企业id，返回关系数据。

## 5.2 安全性

使用Spring Security相关技术栈实现token的用户认证与查询授权。

## 5.3 其他功能

提供用户接口调用审计功能、用户调度频率控制等功能。

# 数仓模型子模块

## 6.1企业画像指标统计规则

详见《企业画像指标规则\_1.0》

## 6.2数据存储介质

### 6.2.1 企业基本信息模块

1. 在dwd基本信息表（dwd\_dim\_company\_info\_d）基础之上进行加工处理，结果存入hbase，rowkey设置为企业id；
2. 企业基本信息表后期如果需要模糊查询，建议存入kudu一份
3. 企业基本信息表保留方式：

hbase每次覆盖更新；

Hive是每日一份分区；

Kudu是每个月保存一份；

### 6.2.2 发票模块（进项发票，销项发票）

1. 统计维度为月季年，其中年是自然年，数据存入kudu，分区模式按照

年的范围划分，联合主键需要带上年度月份字段；

1. 基于dwd模型层的发票数据来进项统计计算后存入kudu一份；

### 6.2.3 交易对手模块

### 6.2.3 三大会计报表模块

1. 三大会计报表汇总表都存放在一个大的汇总表，根据汇总信息点击进入查看明细数据；
2. 明细数据展示为目前报表部门那边的三张会计报表，三张表得区分开，不能合在一张表里；
3. 统计年度为月报，每月存放历史数据；

### 6.2.4 财务模块

1）统计年度为月报，每月存放历史数据；

### 6.2.5 税务模块

1. 统计年度为月报，每月存放历史数据；

### 6.2.6 常用指标宽表

根据常用指标筛选出企业，主要实现接口层“多维检索”功能。

## 6.3关键数据脱敏

# 图数据库子模块

## 7.1 数据同步方式

数仓模型子模块每月通过加工进销项发票数据，获得每家企业的上下游企业分布情况（发票数量、发票金额），并将计算结果以csv的形式存储在hdfs上。

图数据库子模块通过python脚本，解析csv文件，并将数据初始化、倒入到Neo4j中。

## 7.2 图数据存储结构

存储节点有企业类型、代帐公司类型。

存储关系为下游关系类型，关系属性包括发票数量与发票金额。

# 前端展示子模块

# 开发环境的配置

|  |  |
| --- | --- |
| 类别 | 标准配置 |
| 计算机硬件 | 8核，64G内存 |
| 操作系统 | CentOS Linux release 7.8.2003 (Core) |
| 开发语言 | Python3.6，Hivesql，Java 1.8 |
| 数据库 | MySql 5.7 x64,Neo4j 3.5（docker） |
| 网络通信 | 无 |
| 大数据环境 | 线下CDH开发环境 |

# 运行环境的配置

|  |  |  |
| --- | --- | --- |
| 类别 | 标准配置 | 最低配置 |
| 计算机硬件 | 8核，32G内存1台 | 4核，16G内存1台 |
| 软件 | Centos Linux Server 7.5 x64 | Centos Linux Server 7.5 x64 |
| 开发语言 | Python 3.6,Java 1.8,HiveSql | Python 3.6,Java 1.8,HiveSq |
| 数据库 | MySql 5.7 x64,Neo4j 3.5 | MySql 5.7 x64,Neo4j 3.5 |
| 大数据环境 | 线下CDH生产环境 | 线下CDH生产环境 |

# 测试环境的配置

*（1）单元测试与开发环境相同。*

*（2）集成测试环境、系统测试、验收测试环境与运行环境相同。*

Ref:

1. Neo4j 3.5的推荐配置：<https://neo4j.com/docs/operations-manual/3.5/installation/requirements/>