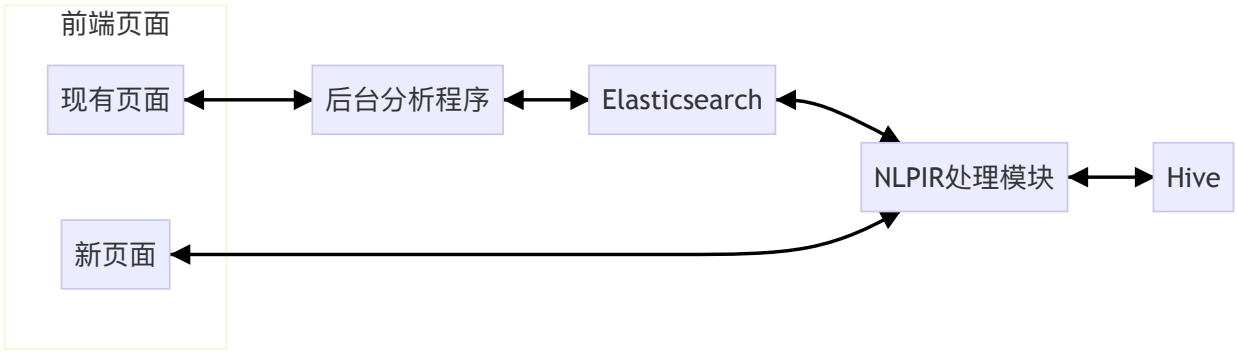
1 现有系统的替换和整合

我们认为现有系统的结构如下:

原始数据存储于 Hive 利用现有的NLP处理模块批量处理数据到Elasticsearch库中,后台分析程序根据处理的结果渲染生成前端页面需要的数据并输出.对于上述的结构,将替换的有NLP模块和部分前端页面:



2 现已经获取的数据和原系统相关信息

现获取到的源系统的相关信息:

1. Elasticsearch 导出的excel形式的字段和少量示例
2. Hive 导出的excel形式的字段和少量示例

现有的信息中的问题:

1. 现有数据字段为excel形式的,和真实数据库中的字段格式有可能存在出入,应该提供原始的字段信息和和现有系统的数据的读权限供我方驻场人员查看.
2. Hive方面的数据接入方式不明,无法在本地进行模拟现有的环境.

3 还需获取的源系统的相关信息

除现有提供的数据外,还需获取的源系统的相关信息和数据用于开发和测试, 分为如下几类:

1. 现有程序的前段页面和后台分析程序,用于进行改造
2. 现有的自然语言处理模块与Elasticsearch和Hive的链接方式和数据连接方式

3. 现有自然语言处理模块中已经存在的适用于业务的关键词,分类体系等业务相关语料,用于提升最终的NLP处理结果

根据上述需要提供的数据,应提供:

1. **现有数据库的直接读权限**,用于分析现有系统中对接的字段意义和字段类型的信息,包括Elasticsearch和Hive
2. 现有程序,包括**前台页面,后台分析程序,自然语言处理部分的程序**的**源代码**和部署方式等**手册**
3. 若无法提供上述源代码,则提供现有运行的系统所在服务器的**远程SSH登录账号**,可以没有任何写权限,可以对现有的程序进行查看等操作即可

4 时间进度

分为两部分: