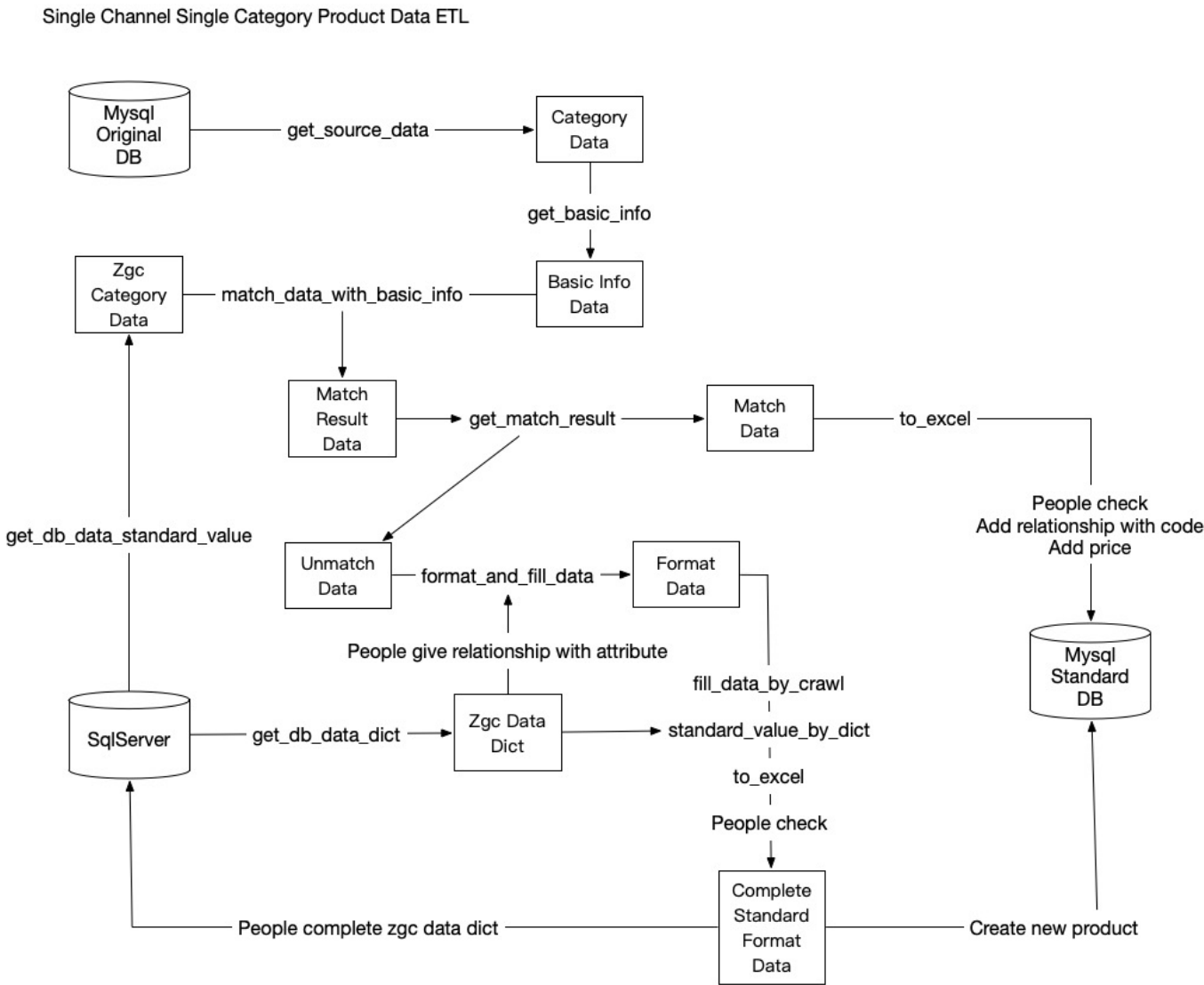


单一渠道单一品类产品数据清洗

流程图



使用步骤

1、get_source_data(channel,category)

输入：渠道名称、品类名称

eg: '天津','复印机'

输出:原始数据(产品数据、产品属性数据、产品价格数据、产品图片数据)

注：参数category用于筛选爬取数据中所需要分析的品类

2、get_basic_info(source_data)

输入：原始数据

输出：产品基础信息

注：需要业务人员确认品牌属性、产品型号属性、价格属性对应的渠道属性名称

3、match_data_with_basic_info(basic_info,category)

输入：产品基础信息、指数库品类名称

输出：匹配结果数据

注：参数category用于筛选指数库中所需的品类数据，请确保该参数值存在于指数库中

4、get_match_result(match_result)

输入：匹配结果数据

输出：匹配上的数据、未匹配上的数据

注：匹配上的数据请使用 df.to_excel()自行导出，并交给业务人员核对并添加编码对应关系，和价格

5、get_db_data_dict(category)

输入：指数库品类名称

输出：指定品类指数库数据字典

注：参数category用于筛选指数库中所需的品类数据，请确保该参数值存在于指数库中

***通过数据字典，获得该品类的标准参数项（已在代码中写出该变量----standard_attribute_list）。请业务人员配合给出该渠道该品类的参数项对应关系，且按信息的可靠程度排序(品牌、型号、价格属性已内置关联，不需要再次对应)**

eg:

```
{'双面器': ['双面功能','product_name'],  
'最大复印尺寸': ['纸张幅面','最大幅面','最大原稿尺寸','product_name'],  
'复印分辨率': ['复印分辨率']}
```

6、format_and_fill_data(source_data,unmatch_data,data_dict)

输入：原始数据、未匹配上的数据、数据字典

输出：格式化且依据对应关系和数据字典补充的数据

注：将参数项对应关系添加至该方法中，覆盖相应位置

7、fill_data_by_crawl(format_data,data_dict,standard_attribute_list)

输入：格式化后的数据、数据字典、标准参数项

输出：爬虫补全后的数据

注：无

8、standard_value_by_dict(data_dict,format_crawled_data)

输入：数据字典、爬虫补全后的数据

输出：结果数据

注：结果数据交由业务人员验证，人工补参、标准化、入库以及补充数据字典