**一． 企业地址识别项目设计流程：**  
 1. 从数据库A（美亚提供）中，获取企业名称（如美亚柏科）；

2. 分别在地图API和招聘网站爬虫，收集其地址数据信息；

3. 再存入数据库B（我这边重新设计该数据库或者在数据库A中添加部分字段）中，来进行地址结果匹配标示；

——在原表中增加字段

4. 最后通过Web或者其他形式（美亚有什么建议），将结果展示出来。

——根据数据库字段，导出CSV格式即可。

因为数据量巨大，在原有企业名称数据的基础上，一一对应进行爬虫收集数据将更有针对性。如果先将企业信息收集下来，再与数据库A中的数据做对比，信息量过大，效果也不好。

——可以先根据企业名称进行采集，下阶段再考虑先采集后对比的方式。

详细细节如下：

**二．企业地址搜集**

1. 通过百度、高德API，输入企业名称、所在城市、所属行业等信息，获得地址相关信息。API调用限制：个人开发者1000-2000次/天，企业开发账户200,000次/天。

1. 对招聘网站，进行爬虫处理，获得地址相关信息。
2. 对企业黄历网站，进行爬虫处理，获得地址相关信息。

目前主要采用1,2两种方法。

**三．数据处理**

通过网上得到的地址数据，和已有的数据进行匹配。需要写一个适度的匹配算法，来判断匹配度。

**同时，希望美亚提供以下支持：**  
 1. 可以提供部分企业数据，如企业名称，城市，行业等。作我们开发测试使用。

1. 开发过程使用的数据库的格式最好能和你们目前存储的企业信息数据库格式一样。在后续项目交接时才可以直接使用。因此美亚能否提供数据库sql结构。

——参考附件中两个工作薄，采集、填充空白字段的内容

**四．运行平台**  
 因为企业地址数据量巨大，是否需要用Hadoop分布式去跑项目，才能支撑一个更加实时的地址识别？

——自行设计

**五．展示**  
 是否需要考虑展示部分？比如在数据库添加一个匹配标识字段。根据之前的匹配去完善填充，更新改字段，并通过web平台实时展示其匹配标示。

——先不考虑展示