甲方名称：

某国企（因保密协议限制无法公开具体信息）

甲方需求：

根据需求定制一采集平台，并进行长期维护。

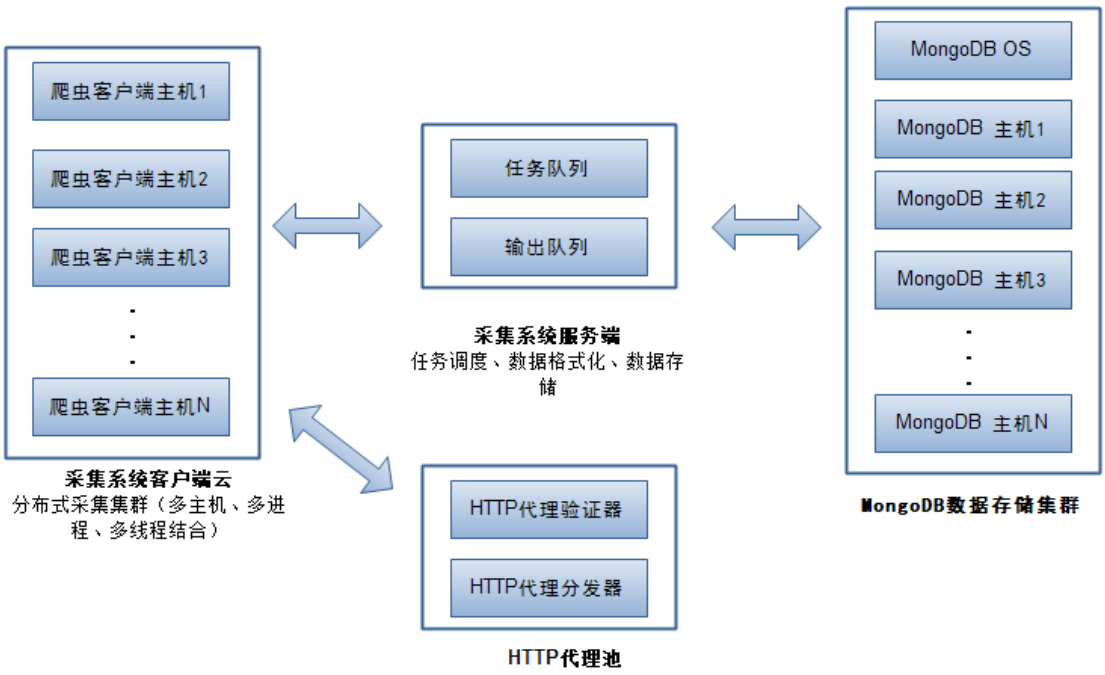
1. 采集目标：去哪儿、携程、卡雅卡、Expedia、天巡等15个机票网站；
2. 从数据库读取待查询线路和日期（未来起飞天数，往返停留天数），分别在各机票网站进行搜索；
3. 提取搜索结果的"航班信息","航程信息","价格信息"存储在数据库中；
4. 采集频率：每日采集一次，保留每次采集结果；
5. 采集量：每天采集机票价格数据超500万条；

技术难点：

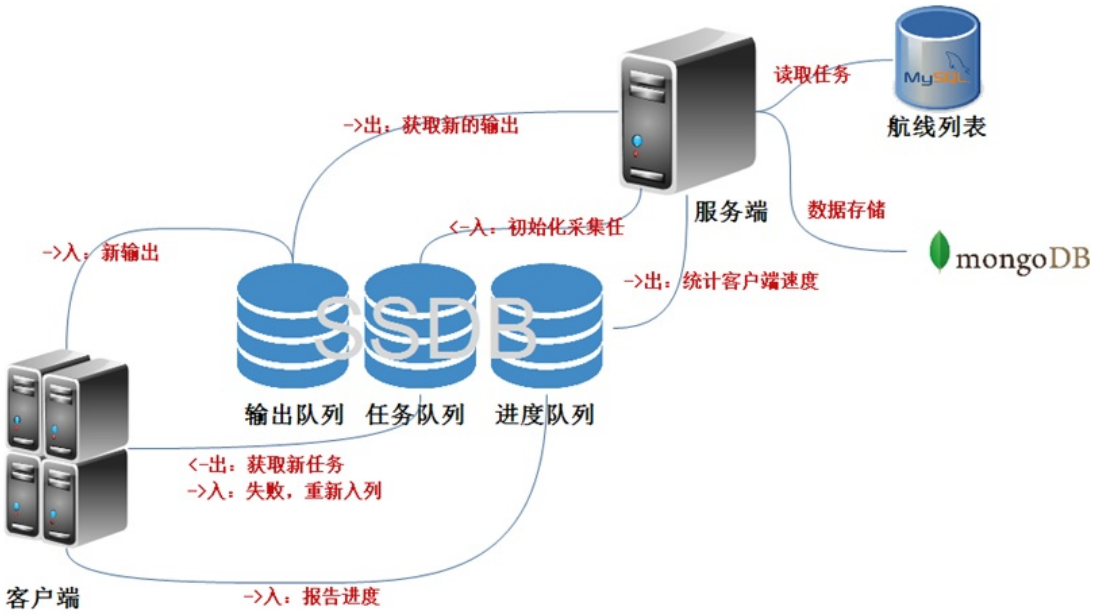
- 航班网站封锁IP非常厉害（特别是卡雅卡、去哪儿网、携程）；
- 部分网站有数据加密策略（例如去哪儿网）；
- 查询任务量非常大，每天查询航线（含日期）数超过20W次；
- 数据输出量比较大，每天输出记录数量超过500W条。

实现方案：

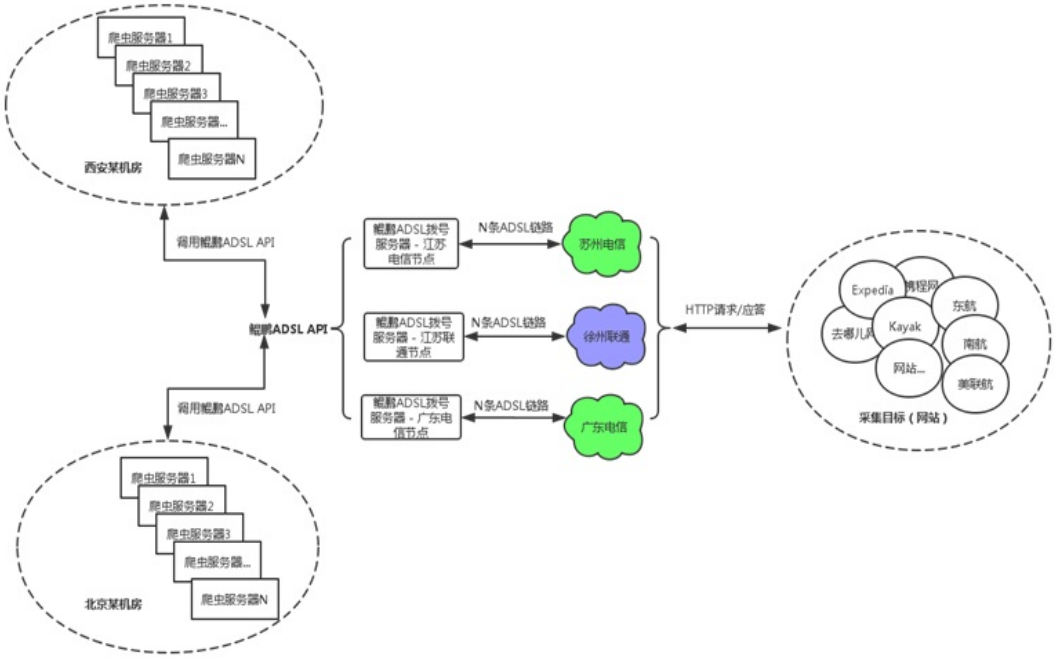
- 通过研究网站的防采集策略，结合鲲鹏ADSL动态IP代理可以有效并稳定持续地采集这些网站的数据；
- 通过Webkit技术有效绕过各种加密策略；
- 采用分布式架构，可以动态扩展采集集群规模（增加采集机器）；
- 采用ActiveMQ作为消息队列，采用MongoDB作为输出数据库；



系统架构图



系统运行流程图



鲲鹏ADSL动态IP代理系统原理图

项目状态：

本项目于2015年7月1日正式开始上线运行，目前仍在稳定进行中。

邮　　编： 710061
联系电话： 029 - 87553281
手　　机： 13571845363 齐工
13389148466 周工
客 服 QQ： 1649677458 或 312602670
邮　　箱： hello@site-digger.com