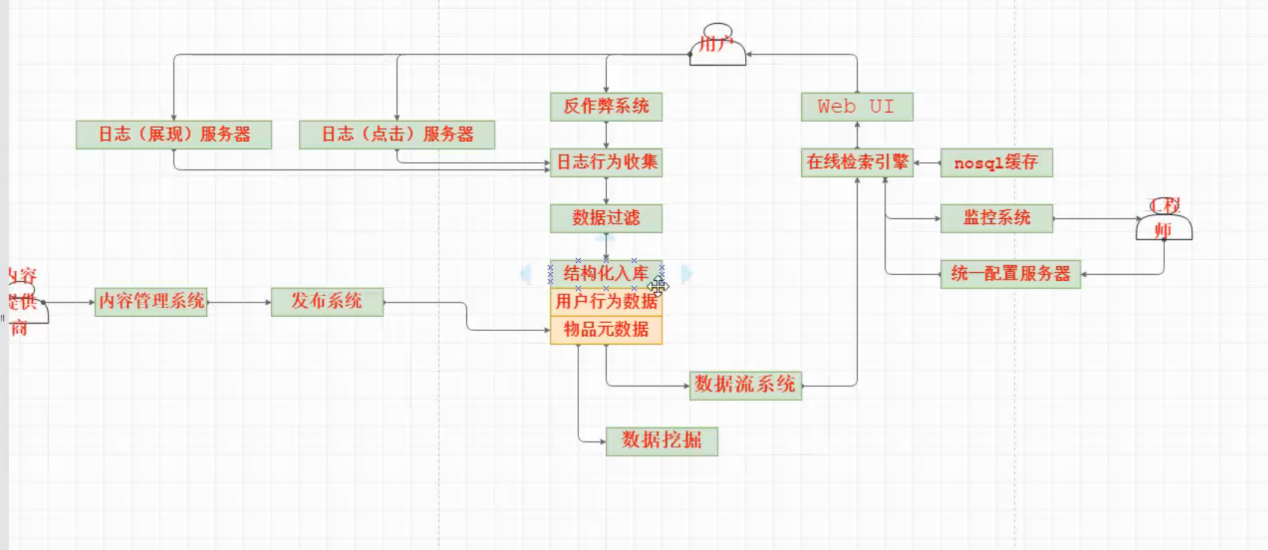
好了，刚说了基本上的一些架构，我们再来看看还这整个架构里面还缺少什么呢？缺少一个数据挖掘平台对吧？（如下图1所示）

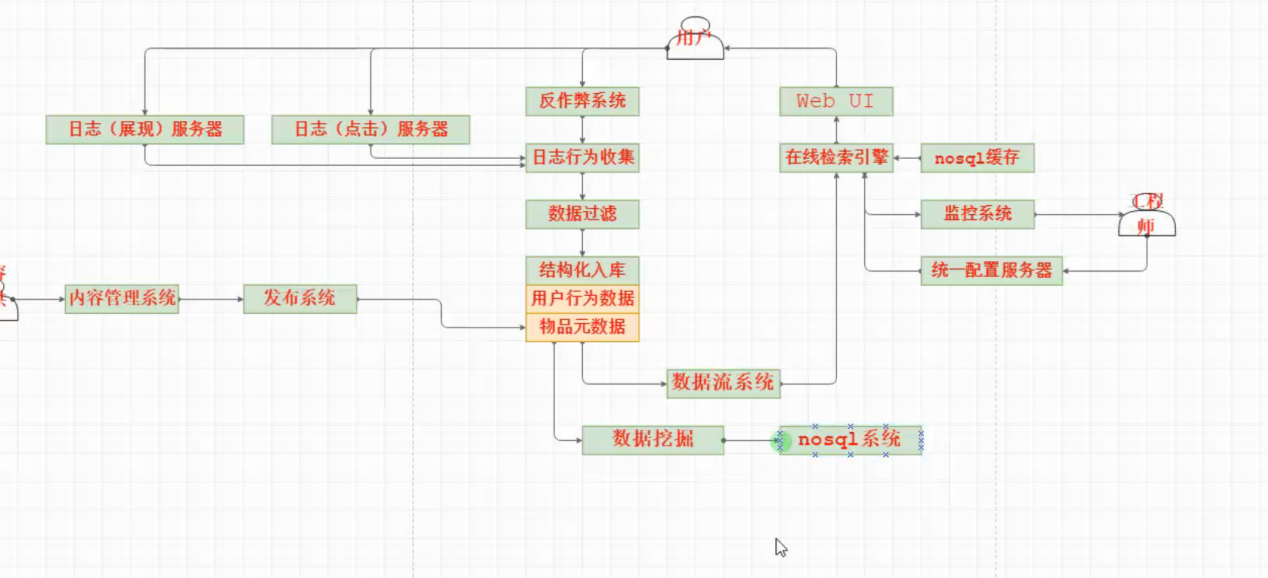
图1：



数据挖掘这里面就会有一些机器学习，算法，需要你的一些离线的分析数据等等，就是对你的大数据真正去做一些个挖掘信息然后供线上使用对吧？那么它这个数据挖掘的数据应该是从哪里引入呢？肯定是从你的结构化入库里面引入过来对吧？比如说这结构化入库就是一个大数据很结构话的数据了，那我们肯定数据挖掘的来源就依赖于数据源对吧？那依赖于结构化入库就可以了，你这结构化入库里面又有用户数据又有物品数据等等很多信息都在这里面，那么我们依赖这个库进行数据挖掘。

那在这个挖掘的过程中，你就可以做一些个算法的提炼，那么这个也是我们后续重点要去做的，那数据挖掘的目的就是得到一些可以供给线上应用的一个有效的数据是不是？那这个数据在真正线上应用的时候也最好给一个缓存系统（Nosql），就是把你这个数据进行缓冲住（如下图2所示）

图2：



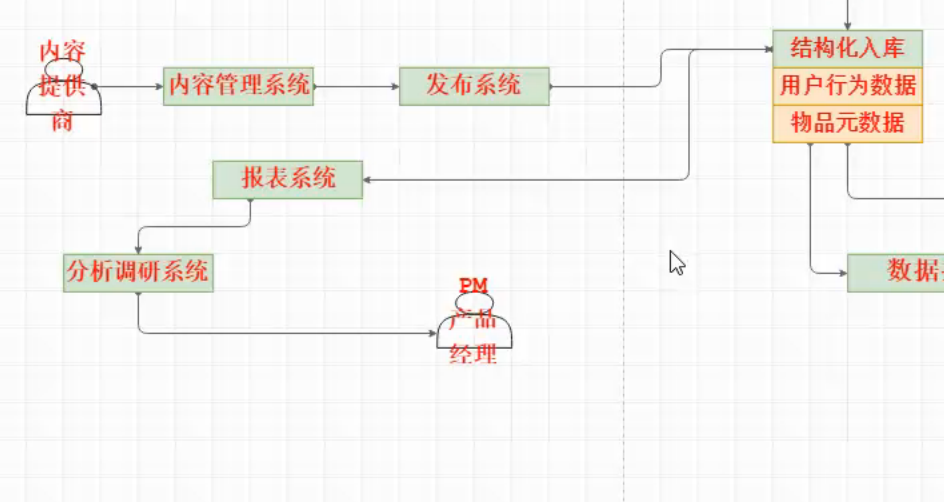
那这个系统需要缓存哪些数据呢？比如说我们要做一个推荐系统对吧？那对于一个我们作为一个用户进来了，那我打开一个交友app的页面，然后它会根据我历史的一些行为能够判断出哪些人是跟我兴趣性格等方面都差不多相同的一些人群，那这个时候相当于是拿着我的这个userid去找很多的那些u1，u2的一些小伙伴是吧？那这些u1和u2等用户是怎么跟我关联起来的呢？都是通过这个数据挖掘系统然后给我通过大数据的分析通过一定的算法并且通过一定的策略然后进行一个数据的提取得到了这些精华部分，把这些精华部分就放到了一个缓存数据库里面。这个数据库就直接供我的线上使用。

比如当前这个webUI页面它就是一个交友网站，那用户来了，这个网站是可以拿到这个用户的id的对吧？然后拿着这个id直接查这个系统看看有哪些跟它相似的用户特征，然后直接给它展示出来就可以了。然后这些数据是怎么匹配出来的呢？就是需要你离线通过一些数据挖掘的一些手段把这些数据给你提前匹配出来。

其实这整个架构的模块别看很简单，其实每一个模块的背后都不是那么简单容易实现的，你要知道这个模块在这个角色里面承担着什么样子的任务，所以具体相当于整个系统大概在你的脑海里面已经有一个整体的印象，那接下来就是要做的能够把系统进行运转起来。

比如有一个产品经理，产品经理是天天提需求，产品调研，用户的一些价值分析和看看哪些金矿还没有挖掘对不对？一旦找到了合适的方向就开始给我们工程师提需求了。比如产品经理是搞分析调研系统，产品经理会通过这个分析调研系统会去对当前的数据不同纬度进行统计，然后看看我想要做这个事情，那这个事情是可行不可行的，是需要一些背后数据的一些验证的对吧？那这个调研系统它很依赖于很多的数据对不对？那这个数据是哪里来的呢？肯定是需要一个数据源对吧？这个数据源就叫一个报表系统（如下图3所示）

图3：

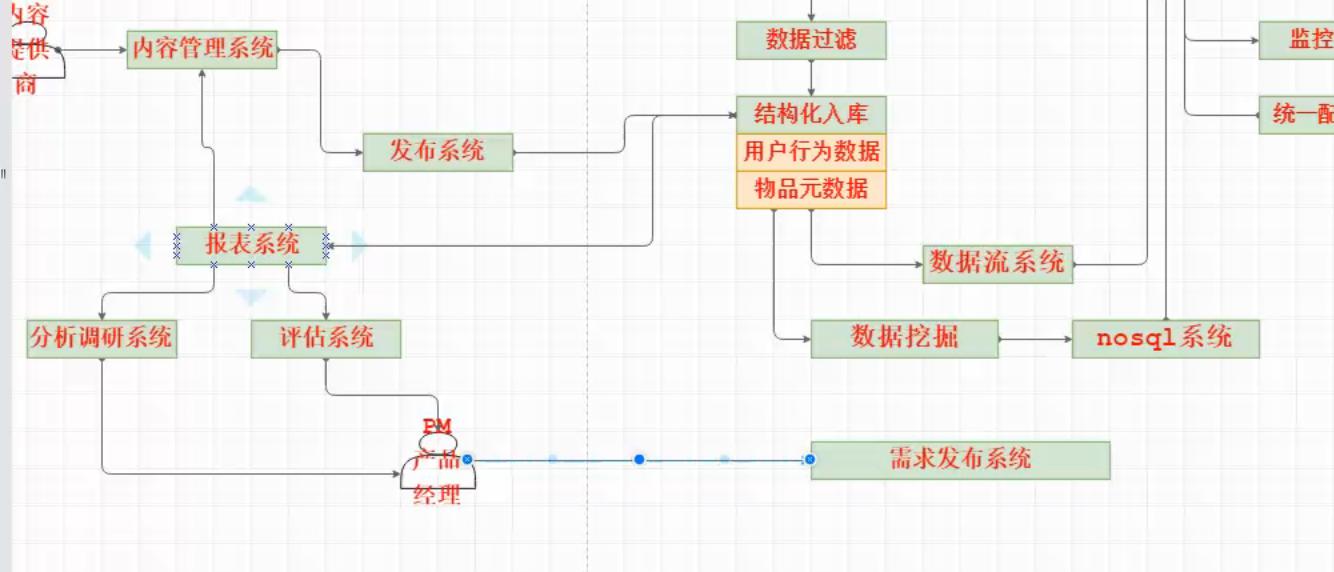


那这个报表系统它的数据是来自于结构化入库这里面的数据的对吧？那么想象我们是一个内容提供商，想象我们是一个老板，然后我们自己去发布一个新的产品然后在内容管理系统里面去写一些产品的描述对吧？你有产品描述的前提是要这个产品真正要去发布出去之后，你应该获得什么样的收益对不对？比如说你这个内容提供商我投的是一双皮鞋，这双皮鞋定价多少多少钱，然后我投进去之后我得知道多少人购买对吧？那我这个皮鞋到底是成本是多少，多少人买，然后最后整个的收益是多少等等这些信息，那这些信息哪里来的呢？这个信息其实在内容管理系统应该有展现是不是？那这个信息也可以通过报表系统来提供给它，那如果要是一个内容提供商它没有这个报表没有这个数据来指导它的话，它其实投出去的商品到底效果好不好，它其实自己也不知道的对不对？那比如说它投了两个产品，然后自己都各投百分之50，其中有个产品在市场反馈来说效果不是很好的，那接下来如果你这个客户继续投这个产品的话必然会导致你的投入产出效率特别底下的是不是？这时候它就完成可以节约成本，把更多的钱投入到更有潜力的方向去或者是探索新的方向去对不对？

那这个时候这个报表系统就应该提供给内容提供商，然后内容提供商优化账户也好，优化它未来的商品投放的策略也好都是一个正向的驱动的意义，这个报表系统其实就是可以用hive去做这样的事情或者是也可以用mysql。

这个产品经理除了做分析调研之外，分析调研是它在做这个产品之前的事情，就是这个产品还没开始，他就提前要做这个事情，如果这个产品已经做好了，然后这个产品的效果怎么样呢？那么产品经理也得知道对吧？所以还有一个评估系统，就是这个产品经理他管理着这个产品到底这个效果怎么样？这个PM产品经理他自己得知道对吧？那如果这个产品效果是不太好，那分析什么原因是吧？如果效果是非常好的话，好了，那接下来再进一步看看有没有改进地方去进一步升级，然后通过一个评估系统进行一个数据的反馈和指导，那它这个评估系统的数据来自于哪里呢？其实也是来自于报表系统（如下图4所示）

图4：



但不管怎么样，产品经理他要最后进行一个需求发布，或者是我发一个邮件给你，能不能在今天下班之前把这个功能实现了是不是？或者是有没有一个平台，在这个平台上然后产品经理会定期的往上面去更新他的一些需求，然后工程师会通过这个平台上去看到一些新的开发需求，那么这个发布系统可大可小。

好了这个就基本上讲完了，刚刚说讲的就是对于一般的一个产品项目来说该包含的内容都已经包含了，但是还是有差异化的产品还是有一些模块没有包含进来的，这个是不管你以后做产品做项目的时候或者是面试的时候起码就会让人觉得你做事情确实是有一个宏观的概念的。

我们正在学的这个hive，都是在讲这个报表这个角色，然后像hbase就是入库这一块，那么mapreduce，spark等这些集成框架的话在日志行为收集，数据流系统和数据挖掘这三个里面都有，其实你学的这个东西它不仅仅是作为某一个角色存在的，只要是适合它的地方就可以去运用，还有zookeeper，这个统一配置服务器怎么通过一个zookeeper去做一个批量分发这个大家估计还记得把？而且很多公司确实是这么用的，就配置方法这一块非常重要，配置方法还有监控这两个怎么通过zookeeper来实现一个监控配置分发这个是很常用的。