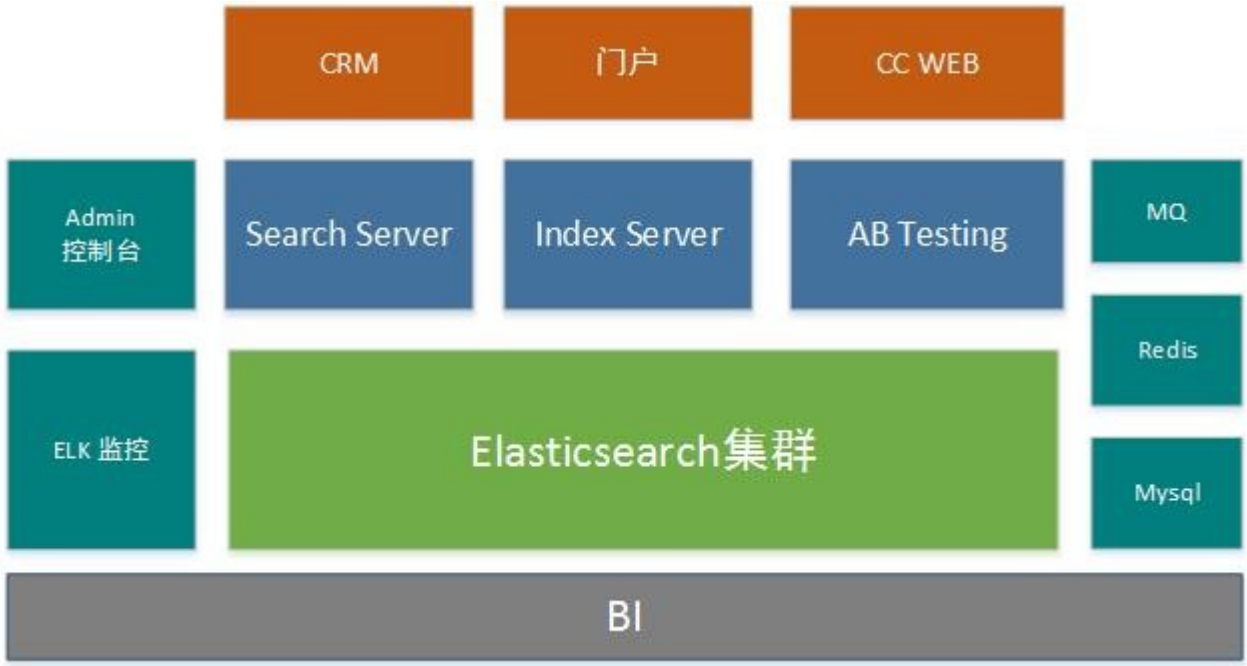
微服务项目之搜索系统实现

一、搜索系统

1.1 简介

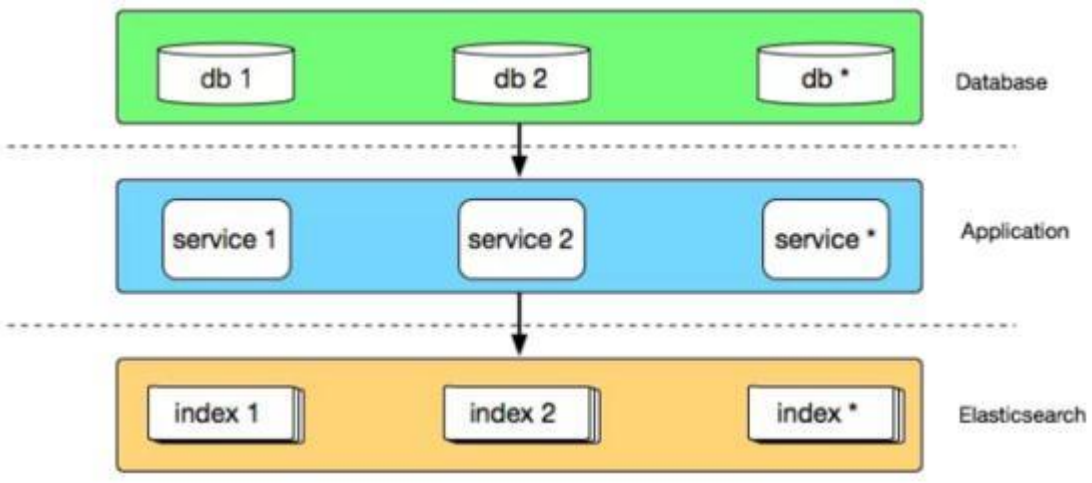
兜儿邦搜索服务底层基于分布式搜索引擎ElasticSearch，ElasticSearch是一个基于Lucene构建的开源，分布式，Restful搜索引擎；能够达到近实时搜索，稳定，可靠，快速响应的要求。



1.2 有赞公司搜索演进

架构 1.0

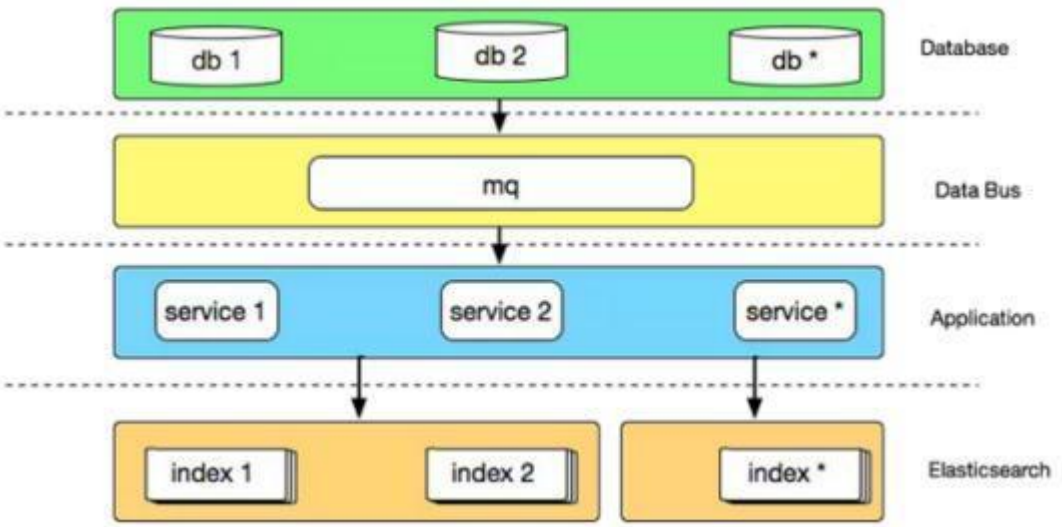
2015 年，运行在生产环境的有赞搜索系统是一个由几台高配虚拟机组成的 Elasticsearch 集群，主要运行商品和粉丝索引，数据通过 Canal 从 DB 同步到 Elasticsearch，大致架构如下图：



通过这种方式，在业务量较小时，可以低成本的快速为不同业务索引创建同步应用，适合业务快速发展时期。 但相对的每个同步程序都是单体应用，不仅与业务库地址耦合，需要适应业务库快速的变化，如迁库、分库分表等，而且多个 Canal 同时订阅同一个库，也会造成数据库性能的下降。 另外 Elasticsearch 集群也没有做物理隔离，有一次促销活动就因为粉丝数据量过于庞大导致 Elasticsearch 进程 Heap 内存耗尽而 OOM，使得集群内全部索引都无法正常工作

架构 2.0

有赞搜索的 2.0 版架构，大致架构如下图：



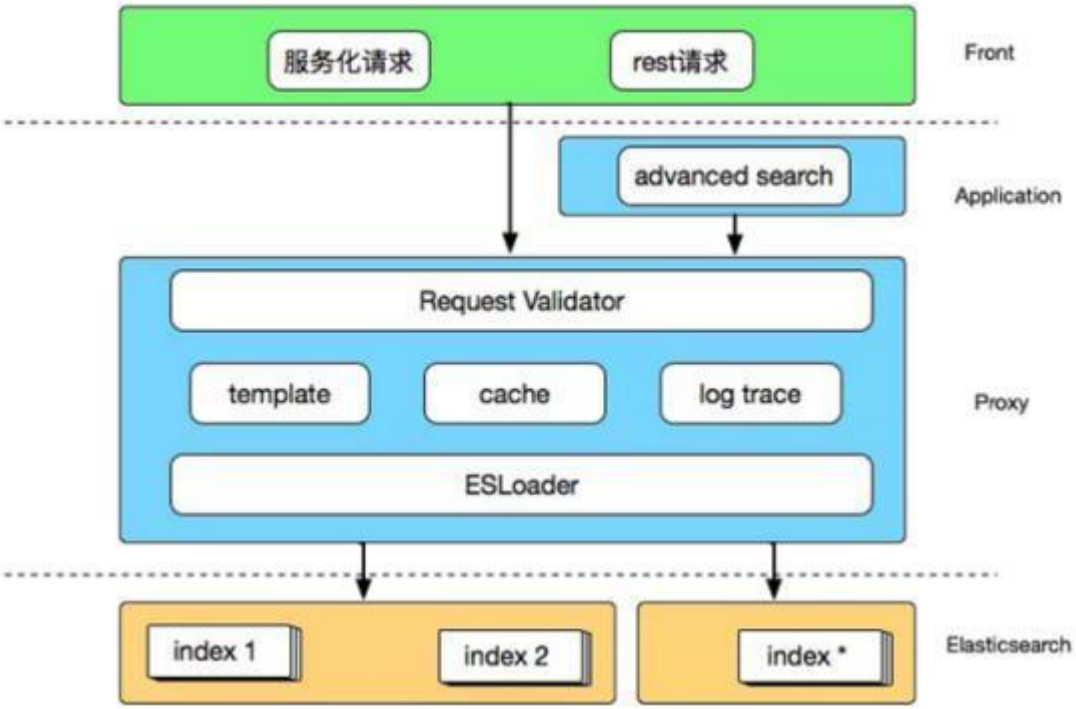
首先数据总线将数据变更消息同步到 MQ，同步应用通过消费 MQ 消息来同步业务库数据，借数据总线实现与业务库的解耦，引入数据总线也可以避免多个 Canal 监听消费同一张表 Binlog 的虚耗。

加入了搜索结果缓存，常规的文本检索查询 Match 每次执行都需要实时计算。

搜索应用和大数据密不可分，除了通过日志分析来挖掘用户行为的更多价值之外，离线计算排序综合得分也是优化搜索应用体验不可缺少的一环。

这样的架构支撑了搜索系统一年多的运行，但是也暴露出了许多问题，首当其冲的是越发高昂的维护成本。除去 Elasticsearch 集群维护和索引本身的配置、字段变更，虽然已经通过数据总线与业务库解耦，但是耦合在同步程序中的业务代码依旧为团队带来了极大的维护负担。 消息队列虽然一定程序上减轻了我们与业务的耦合，但是带来的消息顺序问题也让不熟悉业务数据状态的我们很难处理。 除此之外，流经 Elasticsearch 集群的业务流量对我们来说呈半黑盒状态，可以感知，但不可预测，也因此出现过线上集群被内部大流量错误调用压到 CPU 占满不可服务的故障。

架构 3.0



针对 2.0 时代的问题，我们在 3.0 架构中做了一些针对性调整，列举主要的几点：通过开放接口接收用户调用，与业务代码完全解耦。增加 Proxy 用来对外服务，预处理用户请求并执行必要的流控、缓存等操作。提供管理平台简化索引变更和集群管理，这样的演变让有赞搜索系统逐渐的平台化，已经初具了一个搜索平台的架构。

如果字段数据量比较大，很容易导致 Heap 内存占满引发 Full GC 甚至 OOM。

为了避免重复出现此类问题，我们也提供了定制的可视化查询组件以支持用户浏览数据的需求。

1.3 ES大事件

2013年初，GitHub抛弃了Solr，采取ElasticSearch来做PB级的搜索。“GitHub使用ElasticSearch搜索20TB的数据，包括13亿文件和1300亿行代码”。

维基百科：启动以elasticsearch为基础的核心搜索架构。 SoundCloud：“SoundCloud使用ElasticSearch为1.8亿用户提供即时而精准的音乐搜索服务”。

百度：百度目前广泛使用ElasticSearch作为文本数据分析，采集百度所有服务器上的各类指标数据及用户自定义数据，通过对各种数据进行多维分析展示，辅助定位分析实例异常或业务层面异常。目前覆盖百度内部20多个业务线（包括casio、云分析、网盟、预测、文库、直达号、钱包、风控等），单集群最大100台机器，200个ES节点，每天导入30TB+数据。

新浪，阿里，有赞等著名公司也开始了ES方面的技术研发和实践。

1.4 搜索系统核心功能

实现整体系统的主站搜索（站内搜索）

可以实现商品名称或商品类型的关键字搜索

对外接口：

1、搜索接口

2、搜索日志

1、搜索

2、数据同步

技术选型：

SpringBoot+Spring Cloud

Mybatis+Spring Task+Redis+RabbitMQ

二、搜索系统数据设计

ES索引设计

1、索引

存储搜索的内容（id+名称+类型）

2、索引

搜索记录（id+关键词+ip+时间）

三、搜索系统实现

1、数据哪里来

Mysql/Redis-------ES

基于Spring Task 实现定时任务 时间间隔4小时

cron 0 0 0/4 * * ?

2、如何数据同步性

基于Redis

新增Key

删除Key

修改Key

MQ进行异步通信

3、如何实现数据搜索

Transport

Spring Data Elasticsearch

操作ES服务器的方法：

1、ElasticsearchRepository

2、ElasticsearchCrudRepository

3、PagingAndSortingRepository

4、CrudRepository

5、Repository