90 | 项目实战一：设计实现一个支持各种算法的限流框架（分析）

王争 · 设计模式之美

结束了开源实战，从今天开始我们正式进入项目实战模块。在开源实战中，我带你一块剖析了几个著名的开源项目，比如 Spring、MyBatis、Google Guava 等，剖析了它们背后蕴含的设计思想、原则和模式。

如果说前面讲开源实战是学习别人怎么做，那现在我们讲项目实战就是带你一块做。在这个过程中，我会带你实践之前学过的设计思想、原则和模式，给你展示怎么应用这些理论知识，让你开发出跟前面那些著名开源项目一样优秀的软件。

在项目实战中，我找了三个稍微有点难度的项目：限流框架、幂等框架、灰度发布组件，带你一起来实现。针对每一个项目，我都会从分析、设计、实现这三个部分来讲解。当然，还是那句老话，项目本身的讲解不是重点，重点还是学习它们背后的开发套路。这才是最有价值的部分。

接下来的三节课，我们讲第一个实战项目，限流框架。今天，我们先讲其中的分析环节，介绍项目背景，分析项目需求。

话不多说，让我们正式开始今天的学习吧！

项目背景

我们先来讲下需求诞生的背景。这个背景跟我们下一个实战项目幂等框架也有关系，所以要从很久很久讲起，希望你能耐心看完，不然后面可能会看不懂。

公司成立初期，团队人少。公司集中精力开发一个金融理财产品（我们把这个项目叫做 X 项目）。整个项目只做了简单的前后端分离，后端的所有代码都在一个 GitHub 仓库中，整个后端作为一个应用来部署，没有划分微服务。

遇到了行业风口，公司发展得不错，公司开始招更多人，开发更多的金融产品，比如专注房贷的理财产品、专注供应链的产品、专注消费贷的借款端产品等等。在产品形态上，每个金融产品都做成了独立的 App。

对于不同的金融产品，尽管移动端长得不一样，但是后端的很多功能、代码都是可以复用的。为了快速上线，针对每个应用，公司都成立一个新的团队，然后拷贝 X 项目的代码，在此基础之上修改、添加新的功能。

这样成立新团队，拷贝老代码，改改就能上线一个新产品的开发模式，在一开始很受欢迎。产品上线快，也给公司赢得了竞争上的优势。但时间一长，这样的开发模式暴露出来的问题就越来越多了。而且随着公司的发展，公司也过了急速扩张期，人招得太多，公司开始考虑研发效率问题了。

因为所有的项目的代码都是从 X 项目拷贝来的，多个团队同时维护相似的代码，显然是重复劳动，协作起来也非常麻烦。任何团队发现代码的 bug，都要同步到其他团队做相同的修改。而且，各个团队对代码独立迭代，改得面目全非，即便要添加一个通用的功能，每个团队也都要基于自己的代码再重复开发。

除此之外，公司成立初期，各个方面条件有限，只能招到开发水平一般的员工，而且追求快速上线，所以，X 项目的代码质量很差，结构混乱、命名不规范、到处是临时解决方案、埋了很多坑，在烂代码之上不停地堆砌烂代码，时间长了，代码的可读性越来越差、维护成本越来越高，甚至高过了重新开发的成本。

这个时候该怎么办呢？如果让你出出主意，你有什么好的建议吗？

我们可以把公共的功能、代码抽离出来，形成一个独立的项目，部署成一个公共服务平台。所有金融产品的后端还是参照 MVC 三层架构独立开发，不过，它们只实现自己特有的功能，对于一些公共的功能，通过远程调用公共服务平台提供的接口来实现。

这里提到的公共服务平台，有点类似现在比较火的“中台”或“微服务”。不过，为了减少部署、维护多个微服务的成本，我们把所有公共的功能，放到一个项目中开发，放到一个应用中部署。只不过，我们要未雨绸缪，事先按照领域模型，将代码的模块化做好，等到真的有哪个模块的接口调用过于集中，性能出现瓶颈的时候，我们再把它拆分出来，设计成独立的微服务来开发和部署。

经过这样的拆分之后，我们可以指派一个团队，集中维护公共服务平台的代码。开发一个新的金融产品，也只需要更少的人员来参与，因为他们只需要开发、维护产品特有的功能和代码就可以了。整体上，维护成本降低了。除此之外，公共服务平台的代码集中到了一个团队手里，重构起来不需要协调其他团队和项目，也便于我们重构、改善代码质量。

需求背景

对于公共服务平台来说，接口请求来自很多不同的系统（后面统称为调用方），比如各种金融产品的后端系统。在系统上线一段时间里，我们遇到了很多问题。比如，因为调用方代码 bug、不正确地使用服务（比如启动 Job 来调用接口获取数据）、业务上面的突发流量（比如促销活动），导致来自某个调用方的接口请求数突增，过度争用服务的线程资源，而来自其他调用方的接口请求，因此来不及响应而排队等待，导致接口请求的响应时间大幅增加，甚至出现超时。

为了解决这个问题，你有什么好的建议呢？我先来说说我的。

我们可以开发接口限流功能，限制每个调用方对接口请求的频率。当超过预先设定的访问频率后，我们就触发限流熔断，比如，限制调用方 app-1 对公共服务平台总的接口请求频率不超过 1000 次 / 秒，超过之后的接口请求都会被拒绝。除此之外，为了更加精细化地限流，除了限制每个调用方对公共服务平台总的接口请求频率之外，我们还希望能对单独某个接口的访问频率进行限制，比如，限制 app-1 对接口 /user/query 的访问频率为每秒钟不超过 100 次。

我们希望开发出来的东西有一定的影响力，即便做不到在行业内有影响力，起码也要做到在公司范围内有影响力。所以，从一开始，我们就不想把这个限流功能，做成只有我们项目可用。我们希望把它开发成一个通用的框架，能够应用到各个业务系统中，甚至可以集成到微服务治理平台中。实际上，这也体现了业务开发中要具备的抽象意识、框架意识。我们要善于识别出通用的功能模块，将它抽象成通用的框架、组件、类库等。

需求分析

刚刚我们花了很大篇幅来介绍项目背景和需求背景，接下来，我们再对需求进行更加详细的分析和整理。

前面我们已经讲过一些需求分析的方法，比如画线框图、写用户用例、测试驱动开发等等。这里，我们借助用户用例和测试驱动开发的思想，先去思考，如果框架最终被开发出来之后，它会如何被使用。我一般会找一个框架的应用场景，针对这个场景写一个框架使用的 Demo 程序，这样能够很直观地看到框架长什么样子。知道了框架应该长什么样，就相当于应试教育中确定了考试题目。针对明确的考题去想解决方案，这是我们多年应试教育锻炼之后最擅长做的。

对于限流框架来说，我们来看下它的应用场景。

首先我们需要设置限流规则。为了做到在不修改代码的前提下修改规则，我们一般会把规则放到配置文件中（比如 XML、YAML 配置文件）。在集成了限流框架的应用启动的时候，限流框架会将限流规则，按照事先定义的语法，解析并加载到内存中。我写了一个限流规则的 Demo 配置，如下所示：


```
configs:
- appId: app-1
  limits:
  - api: /v1/user
    limit: 100
  - api: /v1/order
    limit: 50
- appId: app-2
  limits:
  - api: /v1/user
```

```
    limit: 50
  - api: /v1/order
    limit: 50
```

在接收到接口请求之后，应用会将请求发送给限流框架，限流框架会告诉应用，这个接口请求是允许继续处理，还是触发限流熔断。如果我们用代码来将这个过程表示出来的话，就是下面这个 Demo 的样子。如果项目使用的是 Spring 框架，我们可以利用 Spring AOP，把这段限流代码放在统一的切面中，在切面中拦截接口请求，解析出请求对应的调用方 APP ID 和 URL，然后验证是否对此调用方的这个接口请求进行限流。

复制代码

```
String appId = "app-1"; // 调用方APP-ID
String url = "http://www.eudemon.com/v1/user/12345";// 请求url
RateLimiter ratelimiter = new RateLimiter();
boolean passed = ratelimiter.limit(appId, url);
if (passed) {
  // 放行接口请求，继续后续的处理。
} else {
  // 接口请求被限流。
}
```

结合刚刚的 Demo，从使用的角度来说，限流框架主要包含两部分功能：配置限流规则和提供编程接口（RateLimiter 类）验证请求是否被限流。不过，作为通用的框架，除了功能性需求之外，非功能性需求也非常重要，有时候会决定一个框架的成败，比如，框架的易用性、扩展性、灵活性、性能、容错性等。

对于限流框架，我们来看它都有哪些非功能性需求。

易用性方面，我们希望限流规则的配置、编程接口的使用都很简单。我们希望提供各种不同的限流算法，比如基于内存的单机限流算法、基于 Redis 的分布式限流算法，能够让使用者自由选择。除此之外，因为大部分项目都是基于 Spring 开发的，我们还希望限流框架能非常方便地集成到使用 Spring 框架的项目中。

扩展性、灵活性方面，我们希望能够灵活地扩展各种限流算法。同时，我们还希望支持不同格式（JSON、YAML、XML 等格式）、不同数据源（本地文件配置或 Zookeeper 集中配置等）的限流规则的配置方式。

性能方面，因为每个接口请求都要被检查是否限流，这或多或少会增加接口请求的响应时间。而对于响应时间比较敏感的接口服务来说，我们要让限流框架尽可能低延迟，尽可能减少对接口请求本身响应时间的影响。

容错性方面，接入限流框架是为了提高系统的可用性、稳定性，不能因为限流框架的异常，反过来影响到服务本身的可用性。所以，限流框架要有高度的容错性。比如，分布式限流算法依赖集中存储器 Redis。如果 Redis 挂掉了，限流逻辑无法正常运行，这个时候业务接口也要能正常服务才行。

重点回顾

好了，今天的内容到此就讲完了。我们一块来总结回顾一下，你需要重点掌握的内容。

今天，我们主要对限流框架做了大的项目背景、需求背景介绍，以及更加具体的需求分析，明确了要做什么，为下两节课的设计和实现做准备。

从今天的讲解中，不知道你有没有发现，基本的功能需求其实没有多少，但将非功能性需求考虑进去之后，明显就复杂了很多。还是那句老话，**写出能用的代码很简单，写出好用的代码很难。**对于限流框架来说，非功能性需求是设计与实现的难点。怎么做到易用、灵活、可扩展、低延迟、高容错，才是开发的重点，也是我们接下来两节课要讲解的重点。

除此之外，今天我们还实践了一些需求分析的方法，比如画线框图、写用户用例、测试驱动开发等等。针对限流框架，我们借助用户用例和测试驱动开发的思想，先去思考，如果框架最终被开发出来之后，它会如何被使用。针对具体的场景去做分析，更加清晰直观。

课堂讨论

在今天介绍项目背景的时候，我讲了公司遇到的一个开发问题，并提出了解决方案，你也可以留言分享一下，你所在公司或者项目中，遇到过哪些比较头疼的开发问题，又是如何解决的？

欢迎留言和我分享你的想法。如果有收获，也欢迎你把这篇文章分享给你的朋友。

AI智能总结

本文介绍了一个支持各种算法的限流框架的设计与实现。文章首先介绍了项目背景，讲述了公司发展过程中遇到的问题，如代码质量低、维护成本高等。随后，文章详细描述了对公共服务平台的需求背景，包括接口请求频率过高导致的问题，并提出了开发接口限流功能的建议。作者希望将该功能开发成一个通用的框架，能够应用到各个业务系统中，甚至可以集成到微服务治理平台中。整体而言，本文重点介绍了限流框架的需求背景和设计思路，为读者提供了对该项目的整体认识。

文章进一步对限流框架的需求进行详细分析，包括配置限流规则和提供编程接口验证请求是否被限流。此外，还介绍了框架的非功能性需求，如易用性、扩展性、灵活性、性能和容错性。作者强调了非功能性需求对于框架设计与实现的重要性，并提出了需求分析的方法，如用户用例和测试驱动开发。最后，作者鼓励读者分享自己在项目中遇到的开发问题及解决方案。

总的来说，本文通过介绍限流框架的设计与实现，展现了对项目需求的深入分析和解决方案的探讨，为读者提供了对该技术领域的全面了解和思考。

全部留言 (24)

最新　精选

Jxin

2020-05-29

1.大佬就是大佬，言简意赅。几句话就将大部分小公司发展阶段的场景描述的淋漓尽致。

2.我们公司的项目完全符合上诉故事背景描述。但比这个故事背景更糟糕的是。我们过早的做了微服务拆分，而且做拆分的人感觉真的不会写代码。因为拆分出来的微服务包无法用明确的模型来定义。模型和微服务包呈现了多对多的关联现象。与其说做了微服务拆分，不如说硬是把大单体用rpc技术拆成了多个包。因为耦合严重，无论是软件开发还是计算资源分配上，不仅没有帮助，反而还更糟糕了。

3.更恐怖的是，例子中业务扩展做了项目拷贝的动作，我们也做过。只是上面拷贝的是一个项目，我们拷贝的是从交易平台到供应链总共几十近百的“微服务”。

4.解决方案：硬着头皮，咬着牙齿写功能做维护。反正需求倒排期，出bug连带绩效清0，人走加人。

5.结果：线上bug频繁，人员流动频繁，新人上手困难。

6.请教：如何判断继续维护扩展的成本已经高于重写的成本？如何说服领导开始着手重写？在商业模式不景气时，还有必要做这个吗？

作者回复: 1. 这个也没法量化判断，只能侧面上、主观上、感受上去判断。
2. 给领导说下现状，说下重写计划，说下重写好处，然后立个生死状，剩下的就看领导觉悟了
3. 看领导重不重视，如果重视，你重写好了，你就升职加薪，如果领导不重视，那也别做出力不讨好的事情了

共 22 条评论>

60

Geek_e9b8c4
2020-06-06

专栏质量太高！小争哥，下个专栏是什么？？不要停

作者回复: 下一步计划 关注我公众号吧”小争哥“

1

梦倚栏杆
2020-06-03

老师，这个非功能性需求有常规的考虑list吗？需要考虑哪些点。比扩展性，易用性，容错性更细化一点的常规考虑点

作者回复: 常规的我能想到的就这些了，还有一些非常规的，要根据你的项目来

业余爱好者
2020-05-29

既有架构演变历史，又有领域驱动设计。既有需求分析，又有测试驱动开发。信息量不小。

25

Jackey
2020-05-29

做过限流项目，我们的方案是Redis挂了可以降级为本地内存限流，这种可能出现不均衡的问题，但短时间内也可以接受

19

JustRunning
2020-05-29

部门做网关性产品，之前旧逻辑简单基于redis限流，最近因为redis性能导致服务堵塞，还在考虑怎么修复，初步是打算本地预分配，比如1000/s,有10个容器，每个容器分配阈值90/s和离线100/s,达到90根据redis探测状态是否正常，正常采用redis，异常用离线100/s限流，觉得有点土方法，暂时难点是在线扩缩容后怎么处理，因为扩缩容又是外部厂家的pass平台~希望能从老师这里借鉴一下更好的经验😘 😉 😘

共 1 条评论>

10

辣么大
2020-05-29

哇！信息量大，有实战！好期待争哥后续两节的课程！
大的没做过，科研项目用到的机器人操作系统ROS，做高层的任务规划，可以调用ROS已有的低层的路径规划和机械臂运动规划。程序启动参数使用yaml文件配置。程序实现算法的时候，先是脑中有的大概的思路，然后才开始动手实现。当时运动规划还没实现，就先留了一个接口，但整理流程都能跑的下来，做到了模块相对独立，要等着我运动规划的功能实现，这个算法的完成流程到现在都跑不出来。思路就是把解耦，把功能分开。最后放个github链接吧，一年多主要经历都在做这个：https://github.com/fip-lab/PRobPlan

共 1 条评论>

9

业余爱好者
2020-05-29

限流熔断幂等等公共基础功能已经有很多开源组件与框架。重复造轮子，只是为了学习。

8

Cutler
2020-07-06

跨数据库的事务一致性问题，第三方服务与本地服务，mongdb与mysql等场景需要保证最终一致性，目前采用队列，幂等加重试，对业务侵入比较大，正在研发基于saga模型的分布式事务中间件。

3

马以

2020-05-31

期待下节课

2