09（二） | 数据服务难道就是对外提供个API吗?

2020-04-24 郭忆

数据中台实战课 进入课程 >

讲述：郭忆
时长 17:37 大小 16.14M

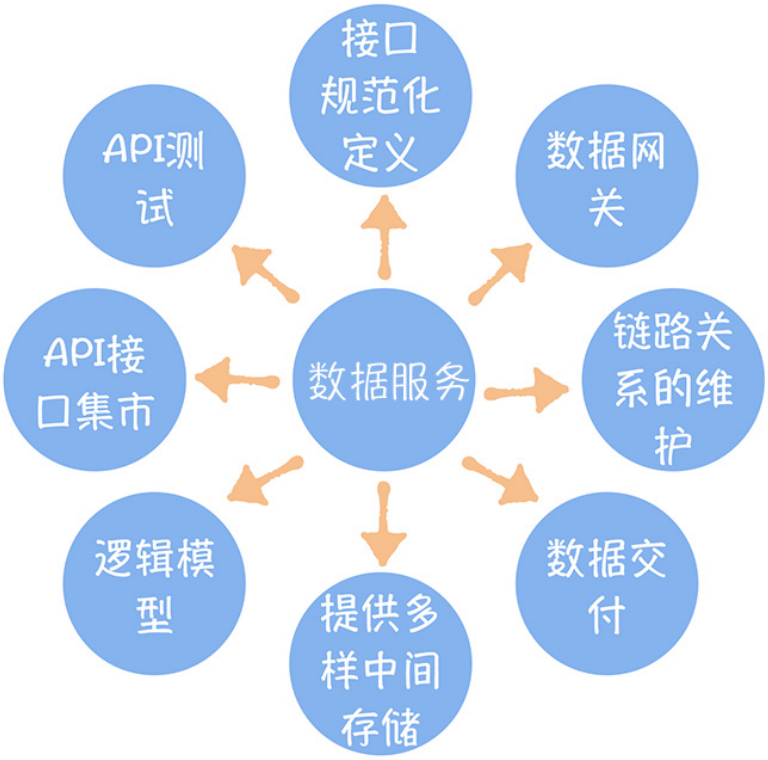
你好，我是郭忆。

在上一讲中，我为你介绍了为什么必须要有数据服务，你可以看到，数据服务在数据建设中发挥着重要的作用。那有的人可能会好奇了，数据服务到底长什么样子呢？ 是不是只对外提供一个 API？ 真的有这么简单吗？接下来，我们就带着这些问题，学习今天的内容。

而我希望你能在学完这部分内容之后，真正掌握数据服务的产品功能设计和系统架构设计。因为这会对你设计一个数据服务，或者选择一个商业化产品，有很大的帮助。

数据服务应该具备的八大功能

我认为，数据服务应该具备八个功能，只有具备这些功能，才能解决我们在上一讲提到的问题。比如，数据接入方式多样，接入效率低；数据和接口没办法共享；不知道数据被哪些应用访问……



那么为了让你更好地理解数据服务的功能，我来讲个小故事。

你肯定去过菜鸟驿站取快递吧？假设有一个很大的菜鸟驿站，里面有很多组货架，每个货架前都有一些工作人员帮助我们取快递，同时也有很多队伍排队。

取快递，要先约定好接口（比如统一使用收货码来取货）。然后，为了保证不同队伍都能取到快递，我们要对每个队伍做一些限流（比如一个队伍一次只能取一个人）。在你取走快递时，驿站会记录是谁取走了哪个快递，方便后续追查。

这段时间，菜鸟驿站服务开始升级，不仅可以取快递，还提供快递送货上门的服务。除此之外，不同种类的快递对应的货架也变得不同，比如生鲜食品，货架是冷藏冰箱，文件、信封，货架就是文件柜。

对于取快递的人来说，如果他买了生鲜，又买了信封，那他要排好几个队伍，肯定不方便。所以，一般来讲，取快递的人最好只在一个队伍排队，而驿站工作人员帮他一次把多个货架的快递都取过来。

可驿站的货架实在是太多了，为了方便每个取快递的小伙伴都能快速找到每个货架以及队伍，驿站提供了一个导览。与此同时，为了不让工作人员出错，驿站的工作人员必须经过严格的测试，才能上岗。

讲完这个故事之后，我们接着回到数据服务的这八大功能上来。在取快递的这个例子中，你可以把数据服务看成是一个菜鸟驿站，工作人员看成是 API 解耦库，货架可以看作是中间存储，快递则可以认为是数据。

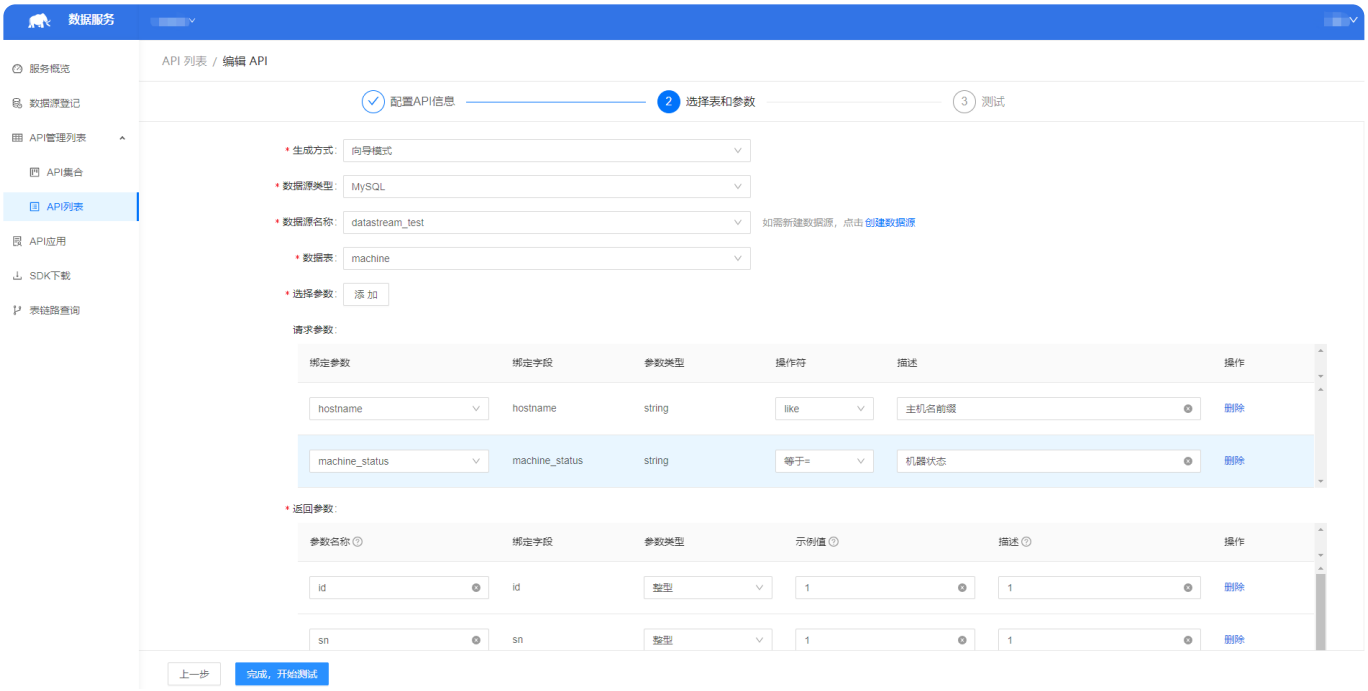
那么对应到八个功能，就是：

- 接口规范化定义，可以看成是取快递约定的收货码，基于统一的收货码取走快递；
- 数据网关，可以看成是我们对每个货架前的队伍进行限流，确保每个队伍都能取走快递；
- 链路关系的维护，可以看作是驿站会记录谁取走了什么快递；
- 数据交付，可以看作驿站同时提供取快递和送货上门服务；
- 提供多样中间存储，可以看成有不同类型的货架；
- 逻辑模型，可以看成是一个工作人员，可以取多个货架的快递；
- API 接口，可以看作是驿站的不同货架的不同队伍导览；
- API 测试，可以看作是驿站工作人员上岗前的测试。

通过这个故事，你是不是已经对数据服务的八个功能有一个形象的感知了？接下来，我们来看看数据服务这八个功能具体包含什么内容。

第一个是接口规范化定义。

接口规范化定义就是取快递时我们约定的取件码。数据服务，对各个数据应用屏蔽了不同的中间存储，提供的是统一的 API。



网易数据服务EasyDS界面示意图

在上图中，我们可以在数据服务上，定义每个 API 接口的输入和输出参数。

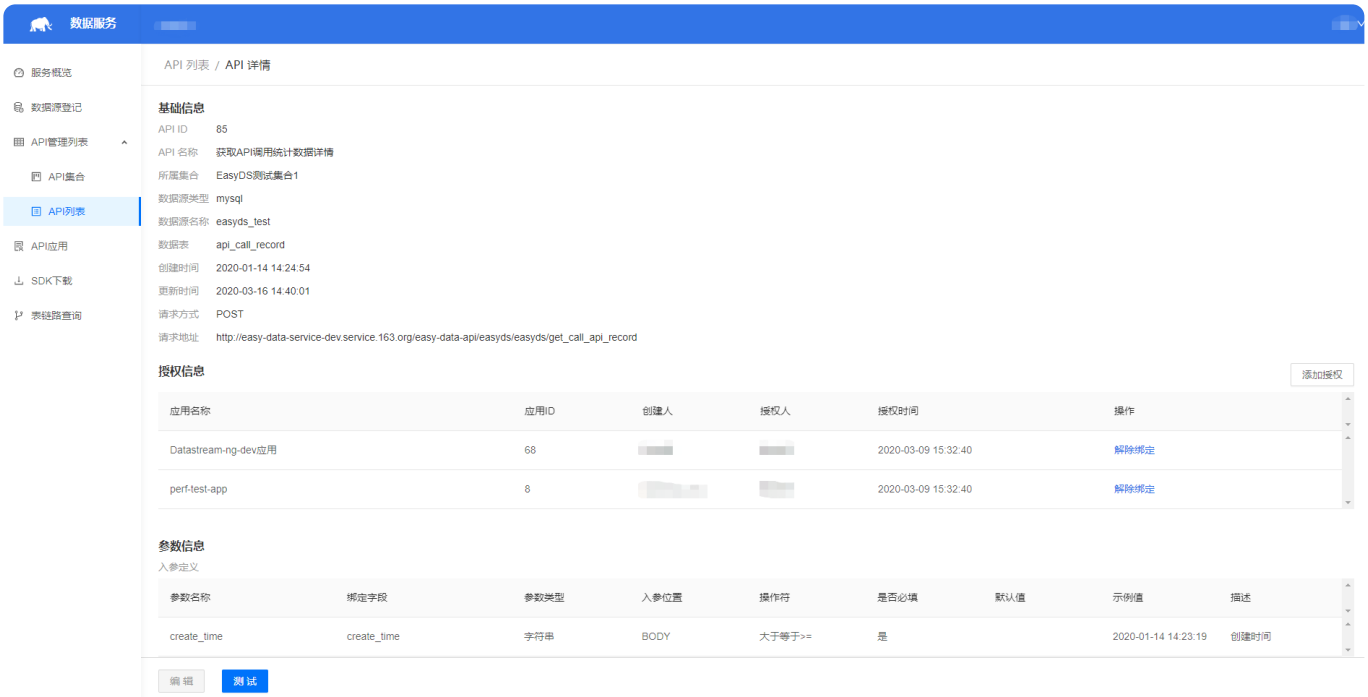
第二，数据网关。

作为网关服务，数据服务必须要具备认证、权限、限流、监控四大功能，这是数据和接口复用的前提。这就跟我们在菜鸟驿站前取快递，要对每个队伍的人进行认证、限流一个道理。我详细介绍一下。

首先是认证，为了解决接口安全的问题，数据服务首先会为每个注册的应用分配一对 accesskey 和 secretkey，应用每次调用 API 接口，都必须携带 acesskey 和 secretkey。

除此之外，对于每个已发布的 API，API 负责人可以对应用进行授权，只有有权限的应用才可以调用该接口。同时，API 接口的负责人可以对应用进行限流（例如限制每秒 QPS 不超过 200），如果超过设定的阈值，就会触发熔断，限制接口的访问频率。

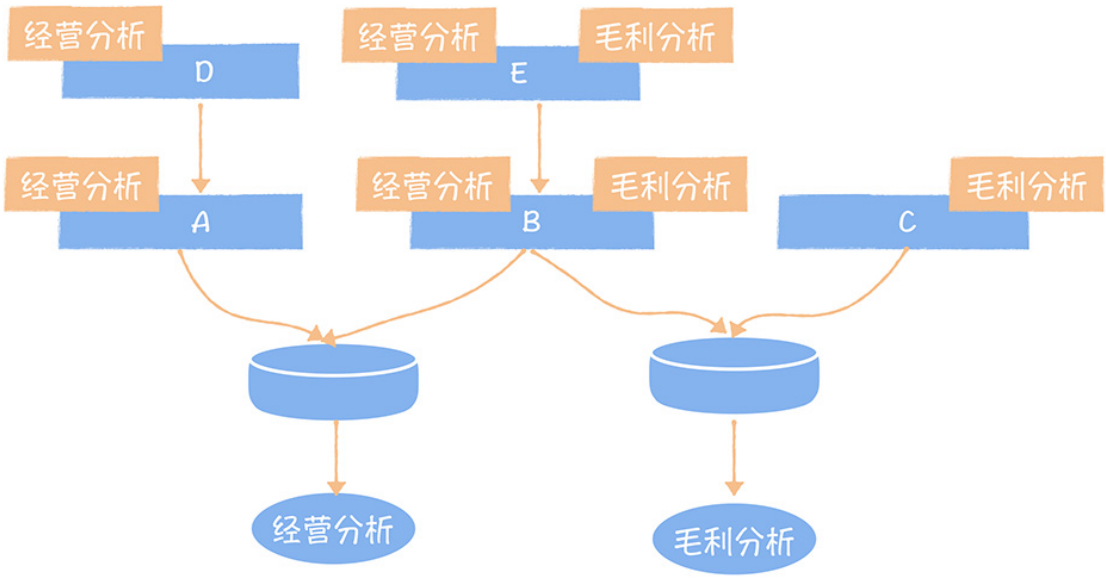
需要你注意的是，对于接口复用来说，限流功能非常必要，否则会造成不同应用之间的相互影响。



应用对接口授权示意图

当然，数据服务还要提供接口相关的监控，比如接口的 90% 的请求响应时间、接口调用次数、失败次数等相关的监控，另外，对于长时间没有调用的 API ，应该予以下线。这样做的好处是防止没用的接口额外占用资源。

第三，全链路打通。



数据服务还必须负责维护数据模型到数据应用的链路关系。

在上图中，经营分析是一个数据应用，甄美丽是数据应用的开发，当她想要访问数据服务中的某个接口获取表 A 和 B 的数据时，她需要向接口的发布者马帅帅申请授予权限。然后经

营分析就可以通过接口获取到数据。

同时，数据服务会把经营分析和表 A 和 B 的访问关系，推送给数据中台的元数据中心。接着元数据中心表 A、B 以及 A 和 B 的上游所有的表（图中 D 和 E）上，就会有经营分析数据应用的标签。

当表 D 的产出任务异常时，马帅帅可以通过元数据中心，快速判断出该任务影响了经营分析数据产品的数据产出。同时，当马帅帅想要下线表 D 时，也可以通过这张表是否有标签，快速判断这个表下游是否还有应用访问。当马帅帅取消 API 接口授权时，元数据中心同时会清理表的相关标签。

需要特别提到的是，一个数据应用往往涉及很多页面，如果我们在影响分析时，只分析到应用，可能粒度还是太粗了，需要到更细级别的页面的粒度，比如一个任务异常，我不光要知道是哪个数据产品，还必须得知道是哪个数据产品的哪个页面。此时，我们在接口授权时，可以标注页面名称。

第四，推和拉的数据交付方式。

相信你听到的数据服务，都是以 API 接口的形式对外提供服务，但是业务实际场景中，光 API 还不够的。我把 API 方式称为拉的方式，而实际业务中同样还需要推的场景。

比如在实时直播场景中，商家需要第一时间获得关于活动的销售数据，此时就需要数据服务具备推的能力，我把它称为数据的送货上门服务。数据服务将数据实时写入到一个 Kafka 中，然后应用通过订阅 Kafka 的 Topic，可以获得实时数据的推送。

第五，利用中间存储，加速数据查询。

数据中台中数据以 Hive 表的形式存在，基于 Hive 或者是 Spark 计算引擎，并不能满足数据产品低延迟，高并发的访问要求，

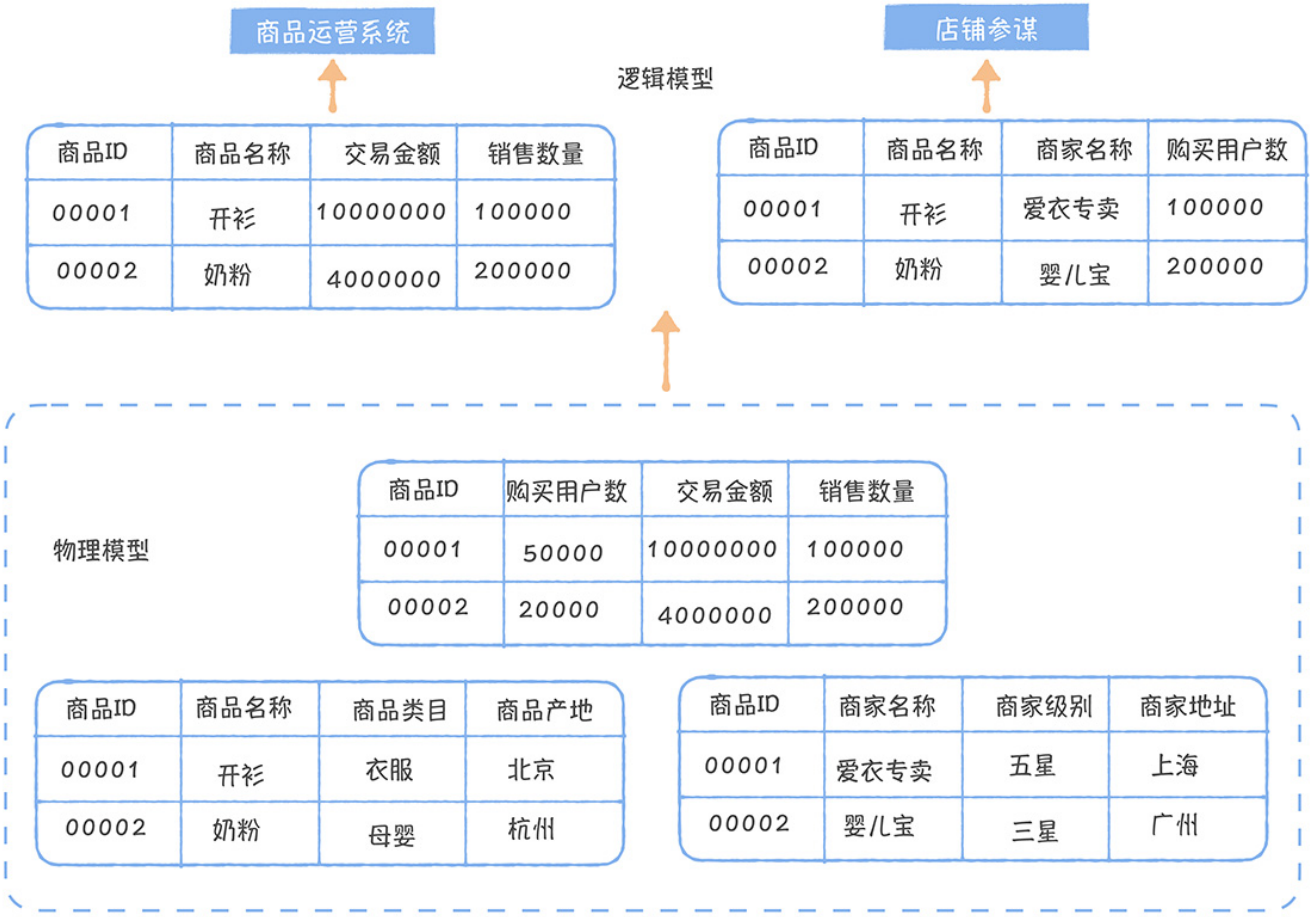
所以，一般做法是将数据从 Hive 表导出到一个中间存储，由中间存储提供实时查询的能力。数据服务需要根据应用场景支持多种中间存储，我列举了一些常用的中间存储以及这些存储适用的场景，希望你能根据实际场景选择适合的中间存储。

中间存储	使用场景
MySQL/Oracle	数据量小，500W记录以内
HBase	数据量大，500W以上记录，基于RowKey 点查，存在冷热明显特征
分布式数据库，例如MyCat	数据量大，都是热数据
GreenPlum	数据量大，多维分析场景
Redis	数据量小，对实时要求比较高

第六，逻辑模型，实现数据的复用。

在前面取快递的场景中，每一个货架一拨工作人员，其实对取快递的人并不友好，所以最好的就是一个人帮我们把所有的快递都取了。这就有点儿类似数据服务中逻辑模型的概念了。我们可以在数据服务中定义逻辑模型，然后基于逻辑模型发布 API，逻辑模型的背后实际是多个物理表，从用户的视角，一个接口就可以访问多张不同的物理表了。

逻辑模型可以类比为数据库中视图的概念，相比于物理模型，逻辑模型只定义了表和字段的映射关系，数据是在查询时动态计算的。逻辑模型可以看作是相同主键的物理模型组成的大宽表。逻辑模型的存在，解决了数据复用的问题，相同的物理模型之上，应用可以根据自己的需求，构建出不同的逻辑模型，每个应用看到不同的列。

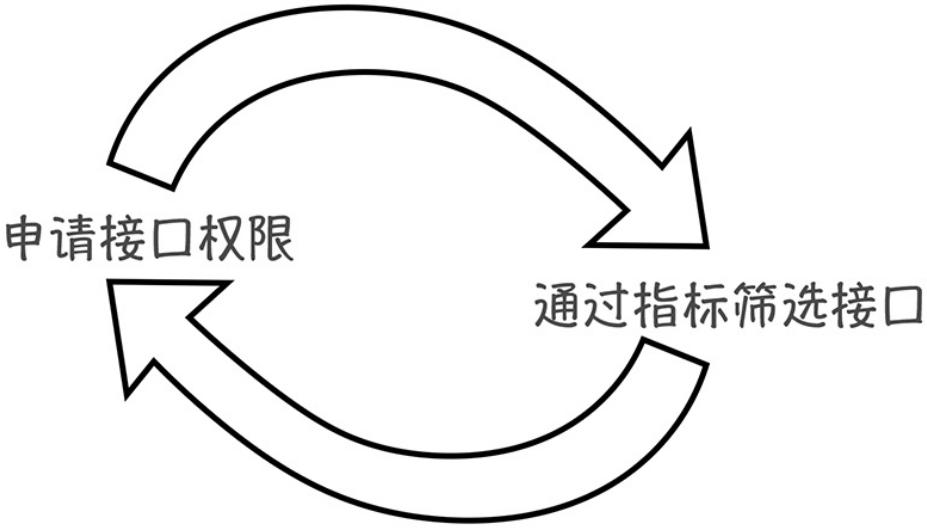


在上面这个例子中，有三个物理模型，但是主键都是商品 ID，针对商品运营系统和店铺参谋，我们可以构建两个不同的逻辑模型，分别从不同的视角看数据，逻辑模型并不实际存在，而是在查询的时候，根据逻辑模型映射的物理模型字段，动态的地将请求拆分给多个物理模型，然后对多个查询结果进行聚合，得到逻辑模型查询的结果。

第七，构建 API 集市，实现接口复用。

为了实现接口的复用，我们需要构建 API 的集市，应用开发者可以直接在 API 集市发现已有的数据接口，直接申请该接口的 API 权限，即可访问该数据，不需要重复开发。

需要特别指出的是，数据服务通过元数据中心，可以获得接口访问的表关联了哪些指标。使用者可以基于指标的组合，筛选接口，这样就可以根据想要的数据，查找可以提供这些数据的接口，形成了一个闭环。



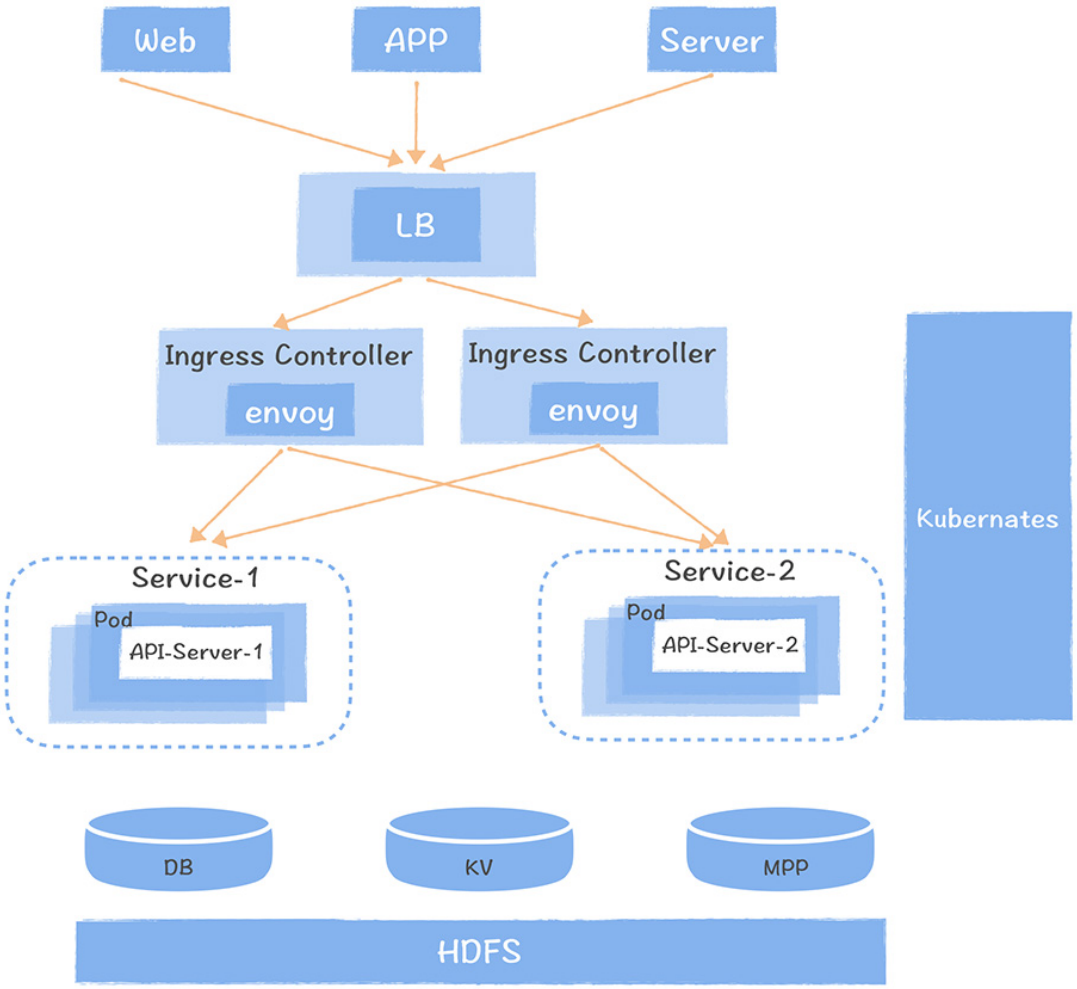
讲了这么多数据服务应该具备的功能，你可能会问，那数据服务应该如何实现呢？ 我们下面就来讲数据服务的架构设计。

数据服务系统架构设计

网易在实现数据服务时，主要采用了云原生、逻辑模型和数据自动导出三个关键设计，关于这部分内容，我希望你能通过学习，在实际工作中可以借鉴我们的方式完成数据服务的设计，或者在选择商业化产品时，给你一个架构选型方面的参考。

云原生

云原生的核心优势在于每个服务至少有两个副本，实现了服务的高可用，同时根据访问量大小，服务的副本数量可以动态调整，基于服务发现，可以实现对客户端透明的弹性伸缩。服务之间基于容器实现了资源隔离，避免了服务之间的相互影响。这些特性非常适用于提供高并发、低延迟，在线数据查询的数据服务。



上图是网易数据服务的部署架构，在这个图中，每个已经发布上线的 API 接口都对应了一个 Kubernates 的 Service，每个 Service 有多个副本的 Pod 组成，每个 API 接口访问后端存储引擎的代码运行在 Pod 对应的容器中，随着 API 接口调用量的变化，Pod 可以动态的创建和销毁。

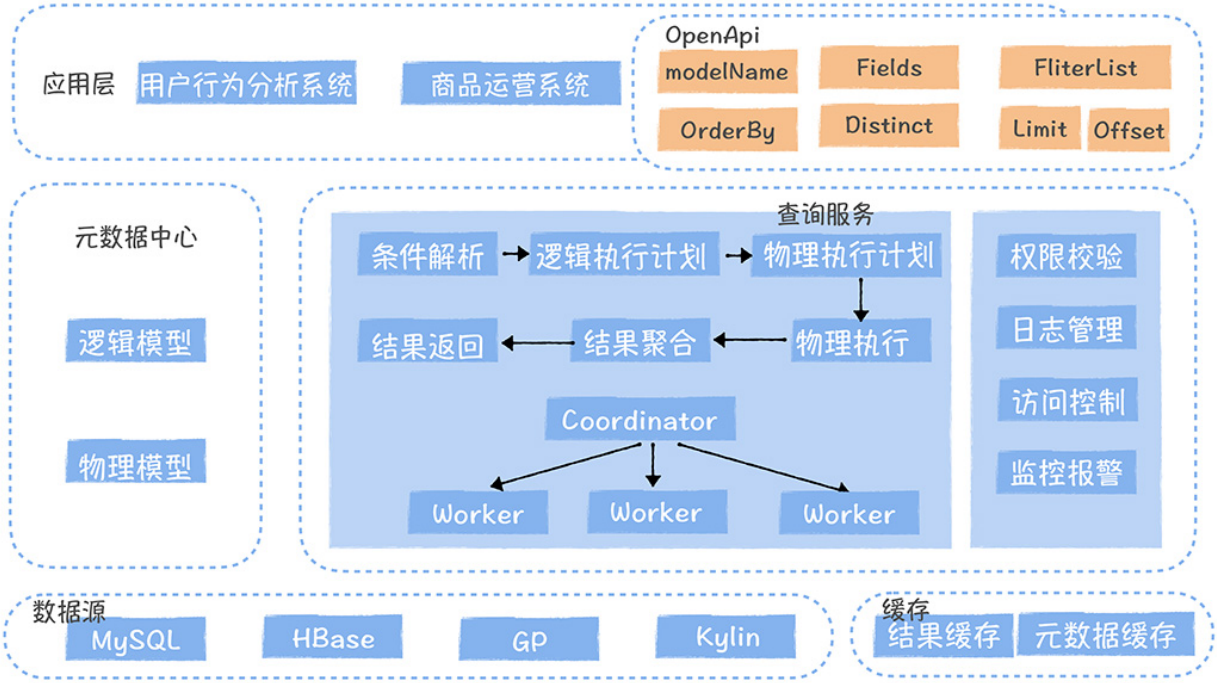
Envoy 是服务网关，可以将 Http 请求负载均衡到 Service 的多个 Pod 上。Ingress Controller 可以查看 Kubernates 中每个 Service 的 Pod 变化，动态地将 Pod IP 写回到 Envoy，从而实现动态的服务发现。前端的 APP，Web 或者是业务系统的 Server 端，通过一个 4 层的负载均衡 LB 接入到 Envoy。

基于云原生的设计，解决了数据服务不同接口之间资源隔离的问题，同时可以基于请求量实现动态的水平扩展。同时借助 Envoy 实现了限流、熔断的功能。你也可以借鉴我们的方案，实现原原生的数据服务设计。

逻辑模型

相较于物理模型，逻辑模型并没有保存实际的数据，而只是包括了逻辑模型和物理模型的映射关系，数据在每次查询时动态生成。逻辑模型的设计，解决了不同接口，对于同一份数据，需要只看到自己需要的数据的需求。

下图是网易数据服务逻辑模型的系统设计图。



接口发布者在数据服务中选择主键相同的多张物理表构建一个逻辑模型，然后基于逻辑模型发布接口。API 服务接到查询请求后，根据逻辑模型和物理模型字段的映射关系，将逻辑执行计划拆解为面向物理模型的物理执行计划，并下发多个物理模型上去执行，最后对执行的结果进行聚合，返回给客户端。

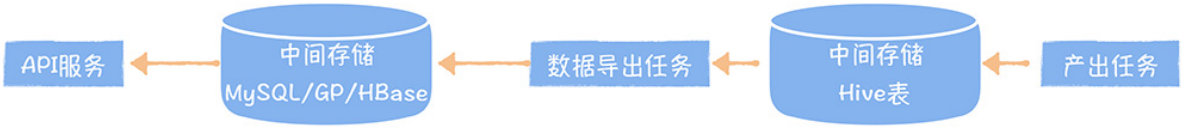
一个逻辑模型关联的物理模型可以分布在不同的查询引擎上，但是这种情况下，考虑性能因素，只支持基于主键的筛选。

数据自动导出

数据服务选择的是数据中台的一张表，然后将数据导出到中间存储中，对外提供 API 。那数据什么时候导出到中间存储中呢？ 要等数据产出完成。

所以在用户选择了一张数据中台的表，定义好表的中间存储后，数据服务会自动生成一个数据导出任务，同时建立到这个数据中台表的产出任务的依赖关系，等到每次调度产出任务结

束，就会触发数据导出服务，将数据导出到中间存储中，此时 API 接口就可以查询到最新的数据。



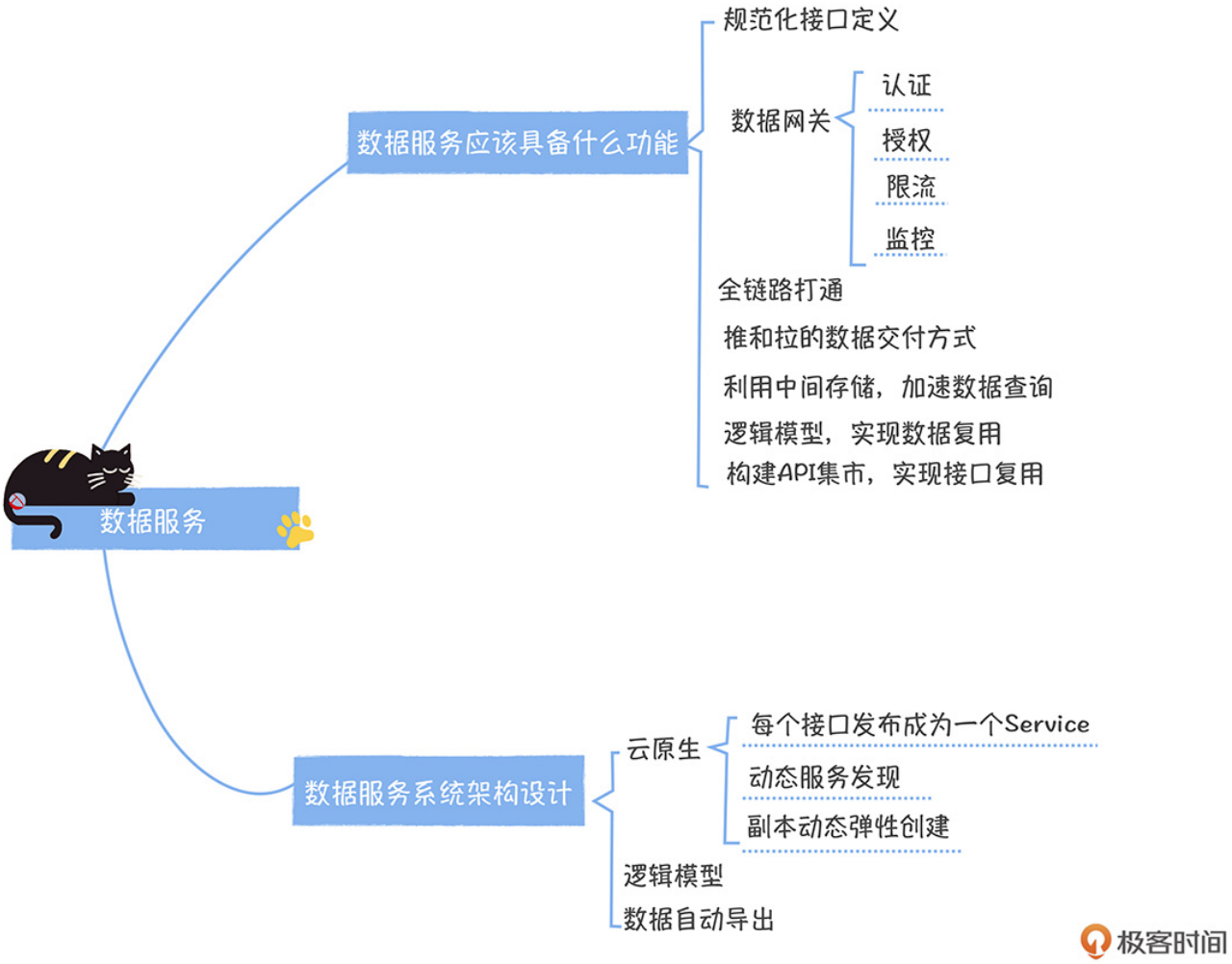
课堂总结

你看，数据服务化不是一个 API 接口这么简单吧，它的背后是数据标准化交付的整套流程。通过这节课，我为你介绍了数据服务的八大关键功能设计和三大系统架构设计。

在最后，我还想强调几个点：

- 数据服务实现了数据中台模型和数据应用的全链路打通，解决了任务异常影响分析和数据下线不知道影响哪些应用的难题；
- 基于相同主键的物理模型，可以构建逻辑模型，逻辑模型解决了数据复用的难题，提高了接口模型的发布效率；
- 数据服务宜采用云原生的设计模式，可以解决服务高可用、弹性伸缩和资源隔离的问题。

数据服务化对于加速数据交付流程，以及数据交付后的运维管理效率有重要作用，也是数据中台关键的组成部分。



思考时间

数据服务要想解决数据被哪些应用访问的问题，就必须确保所有数据应用都必须通过数据服务获取数据中台的数据，那问题来了，如何确保数据服务是数据中台的唯一出口？欢迎在留言区与我互动。

最后感谢你的阅读，如果这节课让你有所收获，也欢迎你将它分享给更多的朋友。

上一篇　09 （一）| 数据服务到底解决了什么问题?

下一篇　10 | 怎么一劳永逸地解决数据安全问题?

精选留言 (6)

写留言

Bill
2020-04-24

到目前为止，几乎和我设想以及实践方式一致。
数据服务这块儿其实可以更简单
- 被动获取：通过1个API解决所有数据请求，只变响应结构即可
- 主动推送：按需按时推送到使用方，做数据增量/全量同步即可。

展开

作者回复: 感谢你的认可，看来你对这个问题也有深入的思考。数据服务，最早我们其实只提供了API的方式，但是其实发现很难满足业务的全部要求，比如实时数据推送的这个场景，你靠API就搞不定。所以我们又做了送货上门的推的服务。

感谢你的阅读，下一次留言区再会~

枕烟客
2020-04-26

要让数据服务成为数据中台的唯一出口，是不是可以换句话说，数据服务能满足数据需求，数据需求能从数据服务中得到需要的数据或服务，这是一个，另外能满足，是不是也要能获得，有这个东西，也得让人家能拿到，不能说，这个东西我们有，就是体验差让人抓狂。

展开

艾伦
2020-04-25

数据服务，更好地打通了数据到业务系统的流转，统一服务建设好，能够更好地实现数据到业务的良性闭环，数据指导业务，业务反哺数据。

展开

gd
2020-04-24

hi，你好！第六，逻辑模型，实现数据的复用，一张表出现2个商品名称。另外，这个实现不同的物理表是服务里面生成查询的视图sql么，如果是不同数据库，是否可以整合，如果可以整合在服务层，是否还需要数据集中？请教下！

展开

阿巍-豆夫
2020-04-24

数据服务太大了。一个数据服务就远比数据治理麻烦的多

> 作者回复: 数据服务，确实技术门槛还挺高的，不过只有数据服务做好，整个数据中台的数据出口才能收口，才能从根本上解决指标管理、全链路血缘建立的问题。
>
> 感谢你的阅读，我们下一次留言区再会~

绍晖
2020-04-24

数据服务就是把部分大不同系统中的表和数据聚合起来对外提供数据服务，中间涉及到逻

辑模型，数据存储中间件，api发布实现等等...

展开

作者回复: 你好，你这里的不同系统，指的是中间存储对吧？ 如果是中间存储是没有问题的。数据服务，其实你可以理解为是数据中台对外提供数据的统一的出口。

感谢你的阅读~