

## 12（一） | 数据研发就只是写代码吗？

2020-05-01 郭忆

数据中台实战课

[进入课程 >](#)




讲述：郭忆

时长 16:09 大小 14.80M



你好，我是郭忆。

到现在，我已经讲了 10 几个数据中台的工具产品，除此之外，我还提到了数据产品、数据架构师、数据开发、应用开发、分析师……多个角色。既然数据中台要用到这么多工具，又涉及这么多角色，如果没有配套的协同流程和规范，那也没办法达到数据中台高效、高质量、低成本的建设目标。来看几件有意思的事儿。

郝有才（数据开发）修改了数据中台一个数据加工任务，变更了产出的数据表字段， 没有通知到下游数据的负责人，结果影响了 10 多个任务，大量数据应用出现异常。这属于比较典型的“协作事故”，咱们再接着看一个跨团队之间协作的问题。

张漂亮（业务系统的服务端开发）今天业务上线，她提交了数据库变更工单，修改了商品交易明细表的商品类型枚举值。但这个升级并没有通知数据部门，结果导致基于商品类型计算的多个指标数值出现错误，严重影响了第二天多个数据产品的数据产出。

这些教训告诉我们，建设数据中台是一项系统性的工程，**你不但要有技术的思维，更要有管理者的视角**。所以接下来，我会带你了解数据中台中，三个最常见的协作流程：数据研发、数据分析、资产管理。**看一下不同角色使用场景化的工具产品，是如何进行高效协作的？**

因为流程协作涉及的料也很多，我会用两讲的时间来讲这部分内容。今天，我们就先从数据研发的场景讲起，如果你是一名普通的数据开发，你肯定很熟悉下面的这些场景。

当然，在学习的过程中，我建议你关注这样几个重点，因为它们对于你理解一个协作流程如何运转非常关键：

一个流程中涉及到了哪些环节？

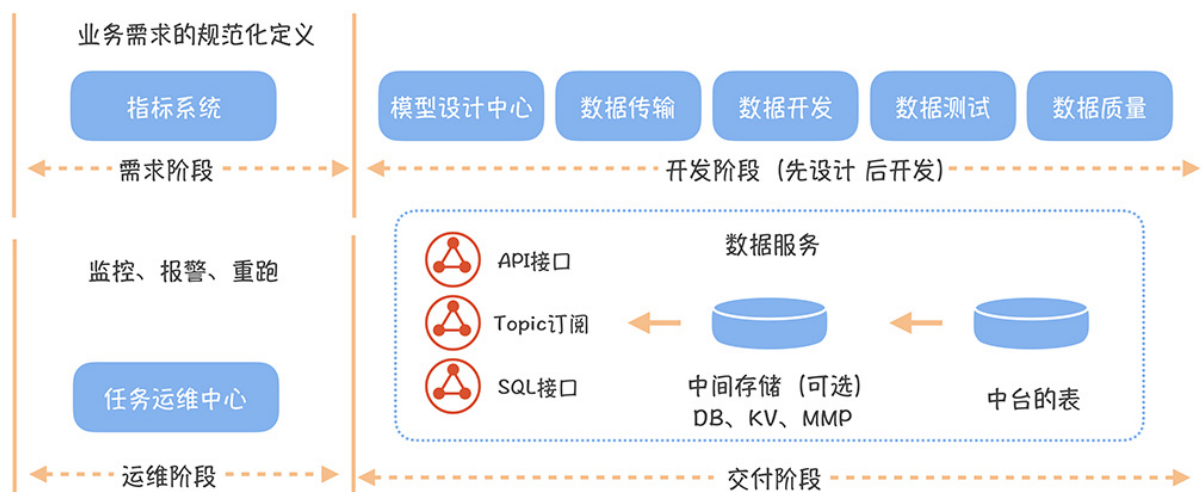
这些环节涉及到哪些角色参与？

承载这个场景的工具产品是什么？

这些环节之间是如何衔接的？

话不多说，开始今天的内容。

也许在很多人的印象中，数据研发就是写代码，其实对大规模、标准化的数据建设来说，这远远不够。在网易，标准的数据研发流程包括四个阶段：需求阶段、开发阶段、交付阶段和运维阶段。每个阶段中又涉及多个环节，如果你缺失了这些环节，就很容易出问题，数据也会因此没办法高效、高质量的交付。



## 需求阶段

需求是数据开发的起点。如果能让后面的流程高效运作，那需求的定义一定要清晰，这样协作者（数据开发、应用开发、数据产品 / 分析师）对需求的理解才能一致。

在数据中台中，数据需求通常是以指标的形式出现的，比如李天真提了个需求（计算每日黑卡会员的消费额），而承载这个场景的产品就是我们 05 讲的指标系统。

那什么时候会提需求？又什么时候会频繁用到指标系统呢？

一般来说，分析师在制作新的报表，数据产品经理在策划新的数据产品时，会提一些新的指标需求，然后就会在指标系统登记指标（包括指标的业务口径，可分析维度、关联的应用、时间周期信息）。这个时候，指标的状态就是待评审状态。

猛犸指标管理系统

配置管理

指标字典

修饰词类型

指标需求

新建需求

Q 请输入需求名称

序号	需求名称	涵盖指标数量	创建人	更新时间	操作															
1	近7天新用户数/新用户销售额	2		2019-09-19 16:17	<a href="#">详情</a> <a href="#">修改</a> <a href="#">删除</a>															
<div>指标列表</div> <table><tr><th>指标名称</th><th>指标口径</th><th>计算周期</th><th>备注</th><th>负责人</th></tr><tr><td>新用户销售额</td><td>新用户的销售额,新用户在单次多次购买,均算做新用户贡献。销售额为含税含运费,去关单回算7天的交易成功金额。</td><td>天</td><td>近7天汇总的新用户产生的销售额</td><td></td></tr><tr><td>新用户数</td><td>首次在考拉平台购买成功的用户数。类目维度下的 新用户,不包含半新用户数。</td><td>天</td><td>近7天汇总的新用户数</td><td></td></tr></table>						指标名称	指标口径	计算周期	备注	负责人	新用户销售额	新用户的销售额,新用户在单次多次购买,均算做新用户贡献。销售额为含税含运费,去关单回算7天的交易成功金额。	天	近7天汇总的新用户产生的销售额		新用户数	首次在考拉平台购买成功的用户数。类目维度下的 新用户,不包含半新用户数。	天	近7天汇总的新用户数	
指标名称	指标口径	计算周期	备注	负责人																
新用户销售额	新用户的销售额,新用户在单次多次购买,均算做新用户贡献。销售额为含税含运费,去关单回算7天的交易成功金额。	天	近7天汇总的新用户产生的销售额																	
新用户数	首次在考拉平台购买成功的用户数。类目维度下的 新用户,不包含半新用户数。	天	近7天汇总的新用户数																	
2		2		2019-09-09 16:27	<a href="#">详情</a> <a href="#">修改</a> <a href="#">删除</a>															
3		1		2019-07-23 14:19	<a href="#">详情</a> <a href="#">修改</a> <a href="#">删除</a>															
4		8		2019-07-23 14:19	<a href="#">详情</a> <a href="#">修改</a> <a href="#">删除</a>															
5		2		2019-07-23 14:19	<a href="#">详情</a> <a href="#">修改</a> <a href="#">删除</a>															

然后，管理指标的数据产品（没有这个角色的，分析师也行）会叫上相关的数据开发、应用开发、提出这个需求的分析师或者数据产品，对指标进行评审：

指标是新指标还是存在的指标；

如果是新指标，那么是原子指标还是派生指标；

确认指标业务口径、计算逻辑和数据来源。

那评审后的结果又是什么呢？

如果是新指标，就在指标系统上录入相关信息，指标状态是待开发状态；

如果是存在的指标，应用开发可以直接找到这个指标所在的表。然后看这个表是否已经有现成的接口可以被直接使用，如果有，就直接申请授权，如果没有，可以基于这张表发布一个新的接口。

## 研发阶段

现在，新指标的状态是待开发状态，接下来就要进入开发阶段。在这个阶段，你要秉持“先设计，后开发”的理念。为啥这么说呢？

因为很多开发都习惯边开发、边设计，想到哪里，代码写到哪里，这其实并不是一个好习惯。这会造成缺少整体的设计，开发过程中经常出现表结构频繁修改，代码返工，整体研发效率不高。

所以说，我们要先做好模型的设计，而承载这个场景的工具产品就是 [🔗06 讲](#) 的模型设计中心。**这里我再强调一下**，数据开发在设计的过程中，可能要用到一些已经存在的数据，这时就要利用数据地图发现已经存在的表，然后理解这些表中，数据的准确含义。

除此之外，在模型设计过程中，要对模型中每个字段关联前面设计好的指标，以及可分析的维度。比如，我们对下图的 account 字段，标记为指标“用户消费金额”；user 标记为“买家维度”。这个标记会把模型和指标建立关联关系，然后把前面设计的指标落实到了表中。

数仓设计平台

数仓建设概览

主题域

表设计工单管理

与我相关

全部

维度

度量

基础字典

创建工单

表名 基础属性及字段

字段配置

添加字段

#	字段名称	字段类型	字段备注	主键	标准化标签	操作
1	account	STRING		<input type="checkbox"/>	维度	删除 上移 下移

添加分区字段

#	分区字段名称	字段类型	字段备注	操作
1	仅支持字母、数字、_	STRING		删除 上移 下移

业务属性

间接关联指标: 请选择指标, 表内不可重复

全部 请输入指标名称 活跃用户数 "主播自我介... 1日留存主播数 1日用户召回率

云音乐 直播 未分组

上一步 提交 保存

模型设计中心，新建模型示意图

到这一步，模型设计还不算完，数据开发还要提交模型上线工单。工单会根据模型所属的主题域，流转 to 对应域的负责人，并通知对应域负责人进行审批。审批通过后，模型会自动发布到生产环境。

数仓设计平台

数仓建设概览

主题域

表设计工单管理

与我相关

全部

维度

度量

基础字典

工单: 100000337 已驳回

库 music\_lplay

表分层 dws

主题归属 直播/流量

表名配置

表名拼接规则 -

建表说明 -

表名称 dws\_music\_livestream\_click\_sd

表描述 直播点击行为轻度汇总表

SQL语句

```
8 anchor_id bigint COMMENT '主播id',
9 live_id bigint COMMENT '直播id',
10 source_type string COMMENT 'app 类型',
11 live_type string COMMENT '直播类型: video live voice live other',
12 alg string COMMENT '算法标识',
13 resource string COMMENT '资源类型',
14 label string COMMENT 'label',
15 cnt bigint COMMENT '次数'
16 )
17 COMMENT '直播点击行为轻度汇总表'
18 PARTITIONED BY (
19 dt string
```

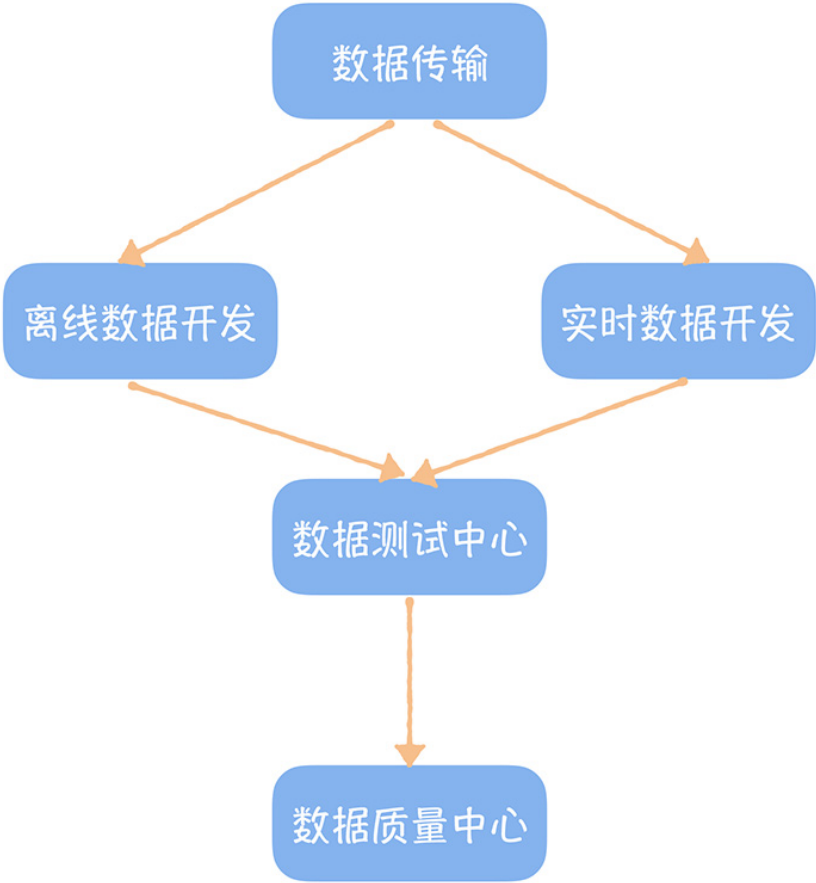
上一步 提交 保存

模型发布上线工单示意图

这里你要注意一下，数据域的负责人一般是数据架构师，他需要检查数据是不是重复建设，要保证自己管理的域下模型设计的相关复用性、完善度、规范性的相关指标。

当然了，除了新建模型之外，已有模型也会存在变更的情况（比如增加一个字段或变更字段枚举值）。这个时候，要根据数据血缘，通知所有依赖这个表的下游任务的负责人，在负责人确认以后，才能进行模型变更。

比如，甄可爱是一名数据开发，她接到需求，完成模型设计之后，就要开始模型的开发了。首先她要把数据从业务系统导入数据中台中，那她第一步就要申请对应数据库的权限，然后在数据传输中心建立数据传输任务，把数据同步过来。



接下来，要清洗和加工数据，那她要在数据开发中心开发数据的 ETL 任务，根据之前模型设计，编写对应任务的代码。

任务代码完成以后，甄可爱要在数据测试中心，验证数据：

一个是进行数据探查，确定新加工的数据是否符合预期；



另外一类是对原有模型的重构，新增字段或者更新部分字段。此时不仅要验证新加工数据的正确性，还要确保原有未修改数据，与修改前是否有改变，我们管它叫数据的比对。

数据测试中心还提供了静态 SQL 代码检查的功能，主要是发现一些使用固定分区，使用测试环境的库，使用笛卡尔积等代码问题，我们把这个过程叫 SQL Scan。在我们的开发规范中，只有通过 SQL Scan 的代码才被允许发布上线。

在数据测试完成后，甄可爱还要在数据质量中心里配置稽核校验规则。目的是对任务产出的数据进行校验，在数据出现问题时，第一时间发现问题，快速地恢复故障。

在开发规范中，主键唯一性监控、表行数绝对值以及波动率监控等属于基础监控，是必须要添加的，另外还需要根据业务过程，添加一些业务规则，比如一个商品只能归属一个类目等。

配置完稽核规则，甄可爱要任务发布上线了。任务发布上线，要设置调度周期，配置任务依赖，设置报警规则以及报警对象，选择提交的队列。

任务发布与模型发布一样，也需要进行审核。首先甄可爱需要发起任务发布上线的工单，然后工单会根据产出表所在域流转 to 对应域负责人贾英俊审批，审批的主要内容：

- 确认任务参数设置是否合理，比如 Spark Executor 分配内存和 CPU 资源；

- 检查任务依赖、报警设置是否正确，核心任务必须要开启循环报警，同时要开启报警上报；

- 重点审核稽核规则是否完备，是否有缺失需要补充。

在审批通过以后，任务就会发布上线，每天就会有数据源源不断的产生了。

到这里，甄可爱就完成了所有模型研发的流程了。你看，虽然是一个模型研发的环节，可涉及这么多的工具产品，还包括了多个审批流程，但是这些工具和流程，都是标准化研发不可或缺的。例如如果不测试，就会导致大量的 BUG 上线，如果没有稽核监控规则配置，就会导致出了 BUG 还不知道，等着被投诉。

而数据研发完，接下来就是数据的交付了，如何让数据快速接入到数据应用中呢？

## 交付阶段

在数据中台之前，其实并不存在单独的交付阶段，因为数据开发加工好数据应用需要的表，他的工作就已经结束了，剩下的就是应用开发的事儿了。应用开发需要把数据导出到应用所属的数据库，然后开发 API 接口，供客户端调用。

数据中台，提出了数据服务化的思想，数据中台暴露的不再直接是数据，而是服务。数据开发不仅需要加工数据，还需要把数据发布成 API 接口或者其他服务形式，提供给业务系统或者数据产品调用，从而形成了单独的数据交付阶段。

数据服务承载了数据交付的整个流程。数据开发，可以直接选择一张数据中台的 Hive 表，然后在数据服务上创建一个数据抽取任务，把数据抽取到中间存储中（中间存储可以是 DB，KV，MPP 等）。这个过程，数据服务会自动根据中台数据的产出时间，在调度系统中创建数据导出任务，建立到产出任务的依赖。

接下来，数据开发可以基于中间存储发布 API 接口，定义输入和输出参数，测试 API 后发布上线。这个时候，数据开发的工作才算完成。

最后，应用开发在数据服务上创建应用，然后申请对该接口的授权，等数据开发审批通过后，就可以直接调用该接口获取数据了。

数据交付完呢，还不算完，接下来数据开发的工作，还需要保证任务的正常运行，这就进入了第四个阶段，运维阶段。

## 运维阶段

承载运维阶段的工具产品主要是任务运维中心。

在这个阶段的第一责任人是任务负责人（一般是这个任务对应的数据开发）。这里有这样几个过程：

数据开发接到报警后，要第一时间认领报警；

任务运维中心提供了报警认领的功能，数据开发点击认领，代表数据开发开始处理这个报警；

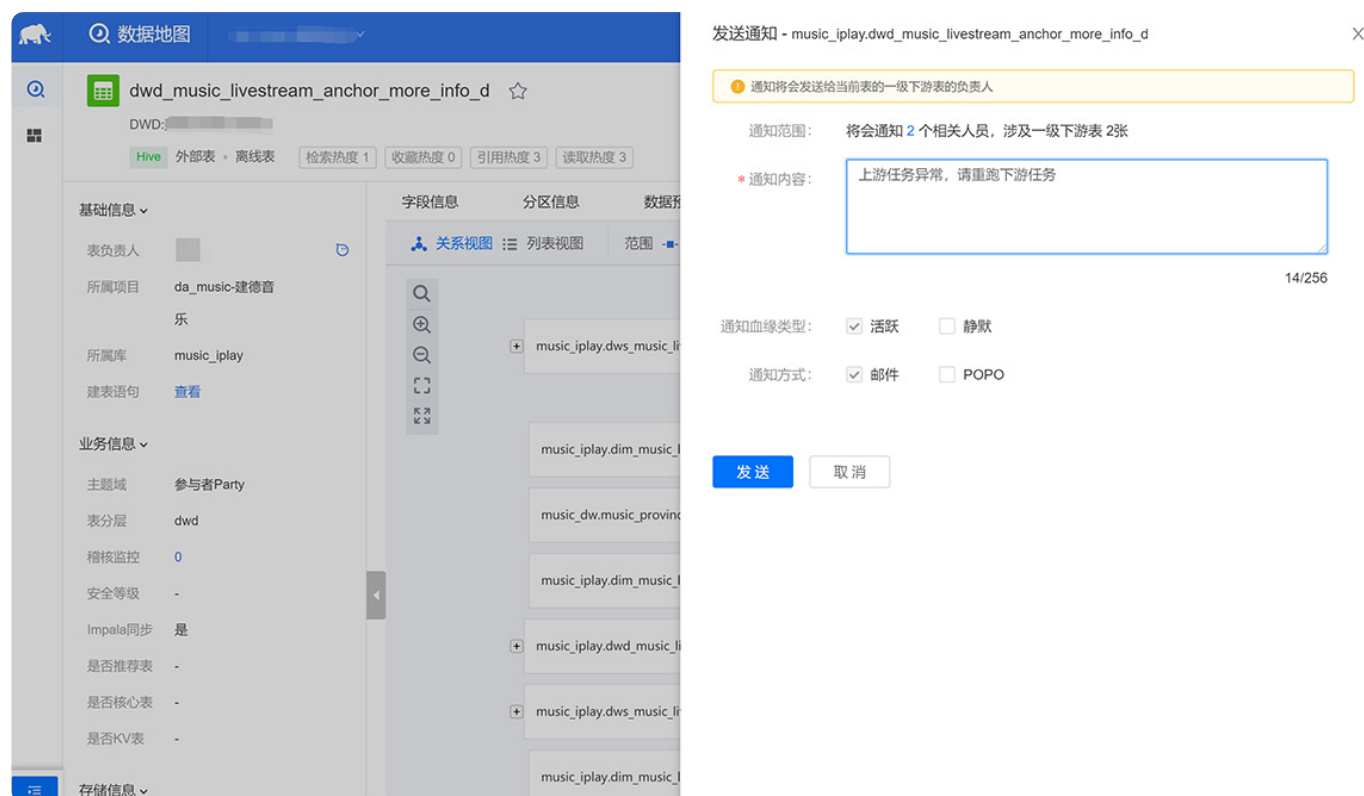


如果报警迟迟没有人认领，任务运维中心会每隔 5 分钟会发起一次电话报警，直到报警认领；

如果报警一直没有认领，系统会在 3 次报警，15 分钟后进行报警的上报，发送给模型所在域的负责人。

这样的机制设计，确保了报警能够在第一时间被响应，我们在实施这项机制前后，报警的平均响应时间，从 2 个小时缩短到 15 分钟内。

那么当数据开发认领报警之后，需要开始排查，首先要确认上游依赖任务稽核规则是否有异常（也就是输入数据是否存在异常）。如果没有异常，数据开发要通过任务运行日志，排查当前任务的问题原因，并进行紧急修复，接下来再重跑该任务，任务重跑完，还要通过数据地图，找到所有依赖该表的下游任务负责人，发送“下游任务需要进行重跑”的通知。



数据地图通知下游任务负责人示意图

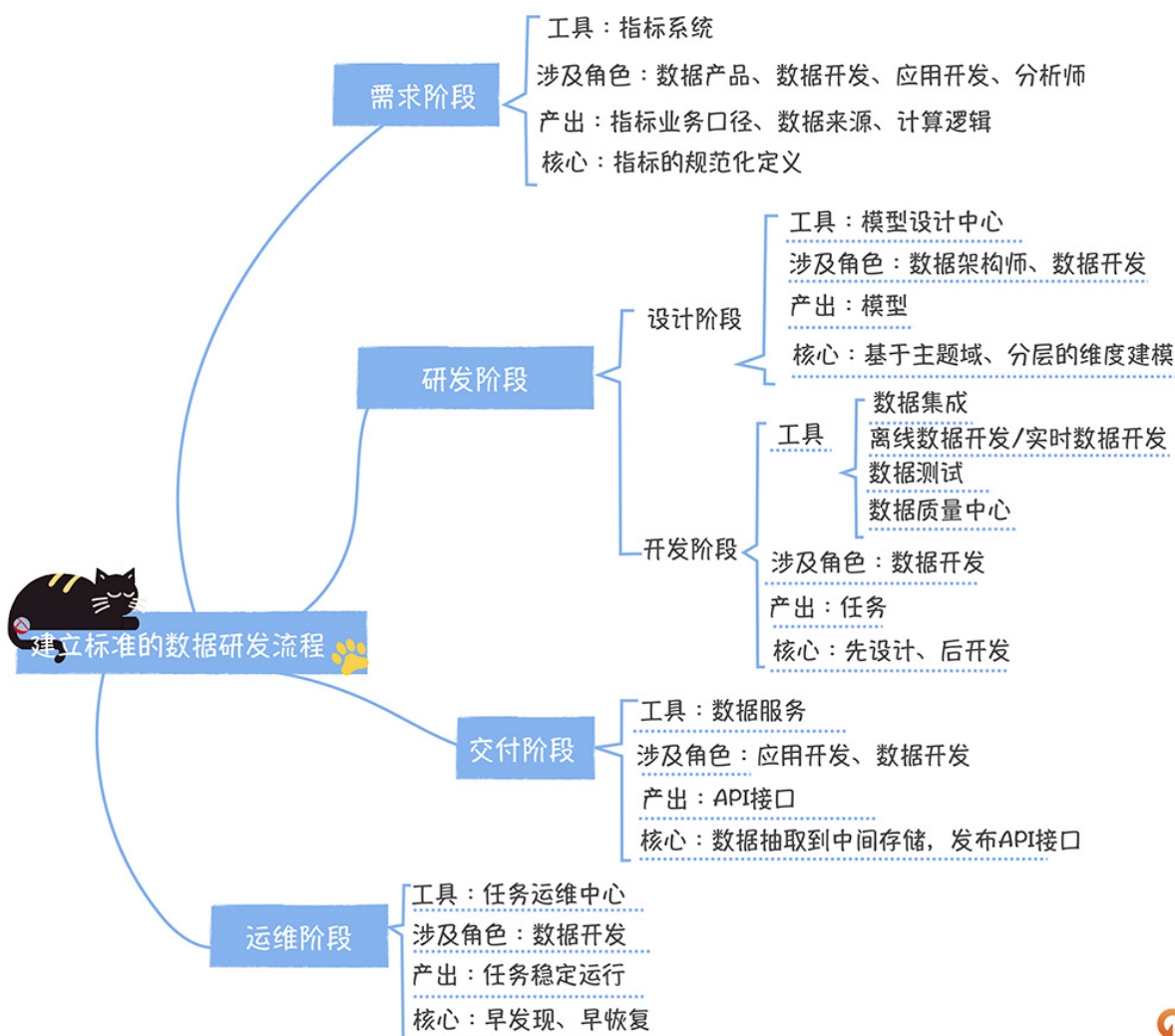
故障恢复完，还要进行复盘，其中重要的事情就是补充稽核规则，确保不再出现犯过的错误。通过这样不断沉淀和记录，数据中台的数据质量就会越来越高，数据质量问题也会减少。

## 课堂总结

你看，数据研发不仅仅只是写代码这么简单吧？在这四个阶段中，你经常容易忽略的是需求阶段和交付阶段，如果需求定义不一致，就很容易导致后面的研发返工，如果没有标准的数据交付流程，就会数据接入慢，同时交付后维护的复杂度会增加。我再强调两个重点：

数据研发的需求是从指标的规范化定义开始，数据产品、数据开发和应用开发要建立一致的指标业务口径、计算逻辑和数据来源，从而才能确保需求被高质量的交付；

数据服务承载了数据标准化交付的功能，通过发布成服务 API 的方式，把数据中台的数据接入到数据产品中。



数据研发好之后，数据就要被使用了，下一节课，我们再以数据使用者的角度以及数据资产管理的视角，带你了解后面两个流程：数据分析流程，带你看一下数据是如何被使用的；然后是资产管理流程，看一下如何有效的实现精细化的资产管理。

## 思考时间

在你日常的数据建设中，遇到过哪些因为流程协作导致的问题呢？欢迎你在留言区与我互动。

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。

## 学习计划

五一计划 📅

# 晒学习姿势 「免费」领课程



【点击】图片，立即参加 >>>

© 版权归极客邦科技所有，未经许可不得传播售卖。页面已增加防盗追踪，如有侵权极客邦将依法追究其法律责任。

上一篇 11 | 数据的台子搭完了，但你还得想好戏该怎么唱

下一篇 12 (二) | 数据被加工后，你还要学会使用和管理数据

## 精选留言 (5)

写留言



leslie

2020-05-03

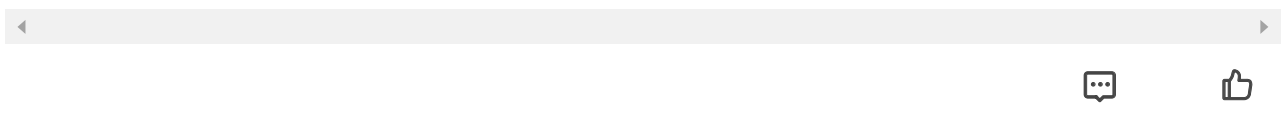
数据中台之中还是涉及了CICD和SRE的东西，看来某些知识其实现在的边界越来越淡了，没有绝对的界限-都是当下。

展开 ∨

作者回复: 对的哦, 看来你也悟到了, 其实数据中台和由微服务构建的技术中台很多技术原理都是相通的, 比如数据服务, 跟微服务中API网关有异曲同工之妙, 这些例子还有很多, 我只能说技术的原理都是相通, 互相借鉴的, 学透了一个, 另外一个就会觉得似曾相识, 当然, 你也可以看看, 数据中台的核心知识点, 是不是可以应用到其他的领域, 比如AI中台等等, 你一定也可以有新的发现。

很多创新都是把一些其他领域实践过方法论应用在一个新的领域, 解决了新领域的某些问题, 所以沉淀这些方法论就很重要, 这也是每个公司为什么在职级晋升答辩时候, 很看重这方面能力的原因。

这里多说了两句, 希望对你有所帮助, 感谢你的留言😊



**aof**

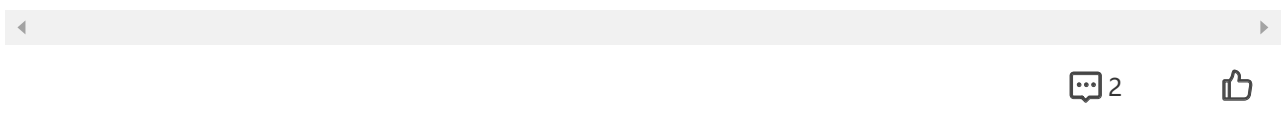
2020-05-02

看了这篇, 真的想去一个真正在做数据的公司!

想想自己公司在做的那些东西, 真是在瞎搞😓😓

展开▼

作者回复: 或者看看是否可以说服自己的老板, 建立一套规范化的数据研发, 管理, 应用流程? 😊



**Sandflass**

2020-05-01

老师五一发课辛苦了, 想问一下老师做思维导图的工具是什么呀? 又可以画导图又可以画表格并关联多个表格, 好赞啊。

展开▼



**吴科**

2020-05-01

五一节, 老师还在发布新课, 赞!

我们业务部门提新需求, 首先去标签系统提需求, 如果是已有的指标就分配相关的权限。如果是新的指标就分配给数据开发人员进行开发。

传统的离线指标, 关键上游系统变更沟通好, 及时通知数据研发部门, 更新元数据管理。数据研发完成后, 根据数据依赖配置好调度, 并设置报警规则。...

展开 ∨



**JohnT3e**

2020-05-01

目前遇到的问题：由于部门建设和人员能力原因，会将ETL纵向切成几个过程，每个过程由不同人去完成，导致整体上缺乏统一考虑，一次排查涉及多个人员，协调起来很费劲。个人认为对于数据开发还是横向切ETL，每个人负责一个或者几个ETL流程方便设计和维护，但同时对数据开发有一定的能力要求。

展开 ∨

