数据全链路

宣讲人：何会会

当前各个使用方的痛点

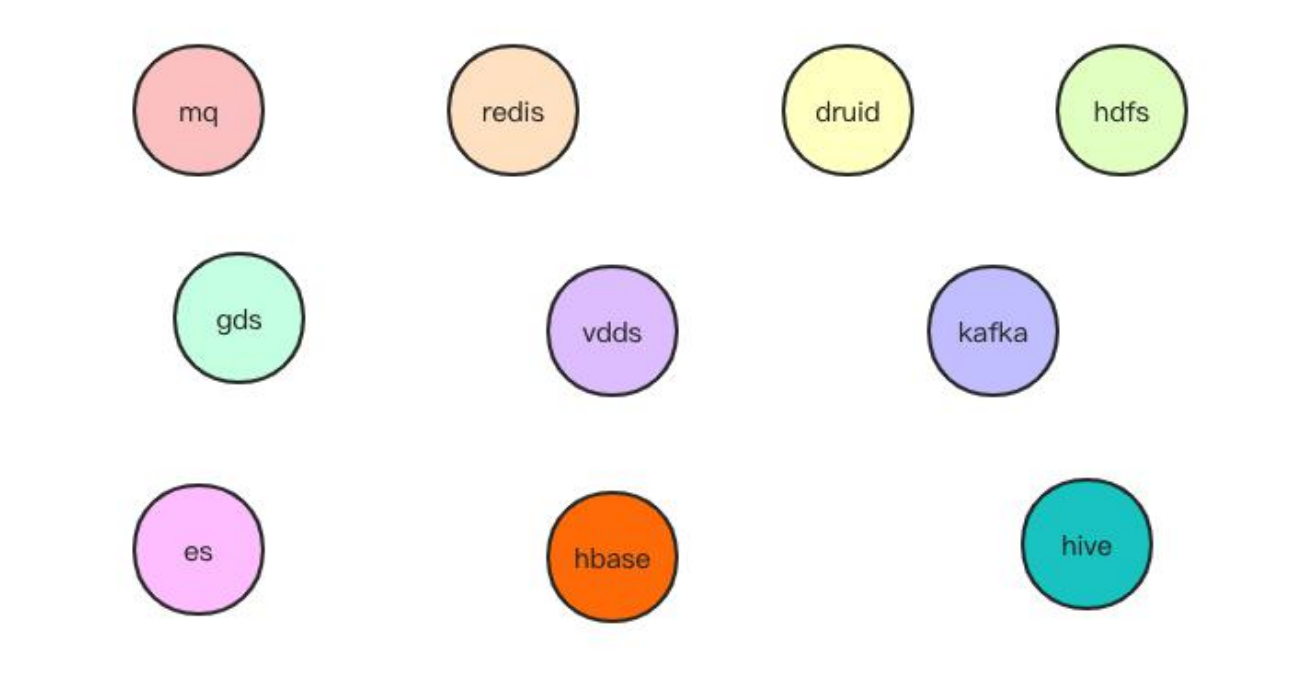
- 数仓：我需要做重构，A表需要下掉，需要使用B表,有哪些下游任务需要切呢？
- BI：我开发的这个报表到底还有没有人看啊，要不要天天关注啊？
- 算法：为啥我的模型天天分配不到资源啊，我们团队的quota不够用啊？
- 我这个任务失败了，影响了多少下游任务啊？
- 这业务下线了，想释放db资源，下游相关的大数据任务和线上业务哪些任务需要搞掉啊？
- 集群管理者，最近集群存储和计算资源有点紧张，需要下批任务，看哪些任务可以下？
- 数据开发：这个表是谁负责的，哪个任务产出的，最近产出有点问题，该找谁帮我看看啊？
- 我这个数据更新了，需要更新下游数据，我要重跑哪个任务？
- vss任务下线了，为啥半夜还给我告警电话说我check任务产出情况，到底能不能下线？

目录

- 数据和任务、全链路的概念及意义
- 如何使用数据全链路快速解决问题

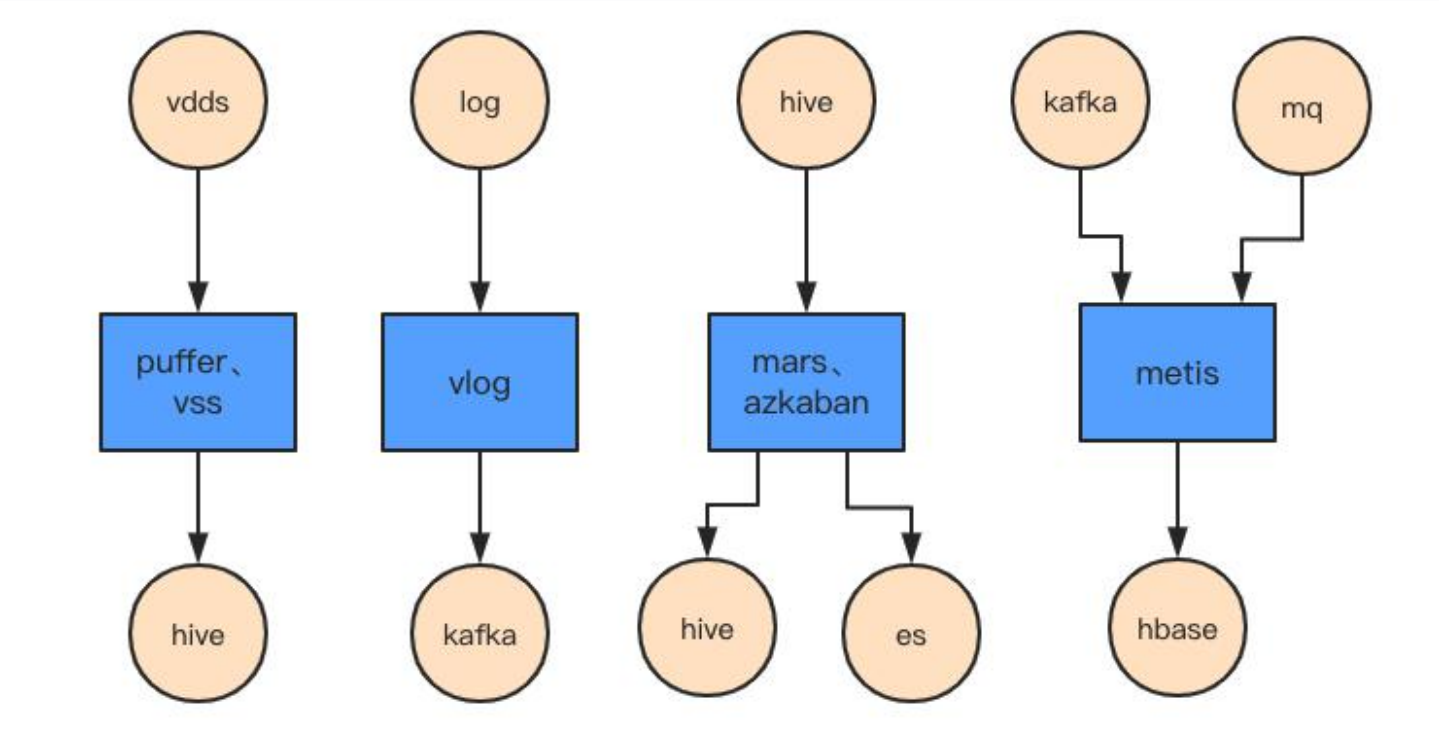
数据的概念和意义

- mq、hbase、es、hive、hdfs、kafka、druid、gds、redis、vdds
- 数据从来不是单独存在的，数据变的有意义在于数据之间的流转、加工处理，最终变成你想要的数据模型和格式
- 资源的唯一标识符：资源类型://所在集群/集群内唯一标识



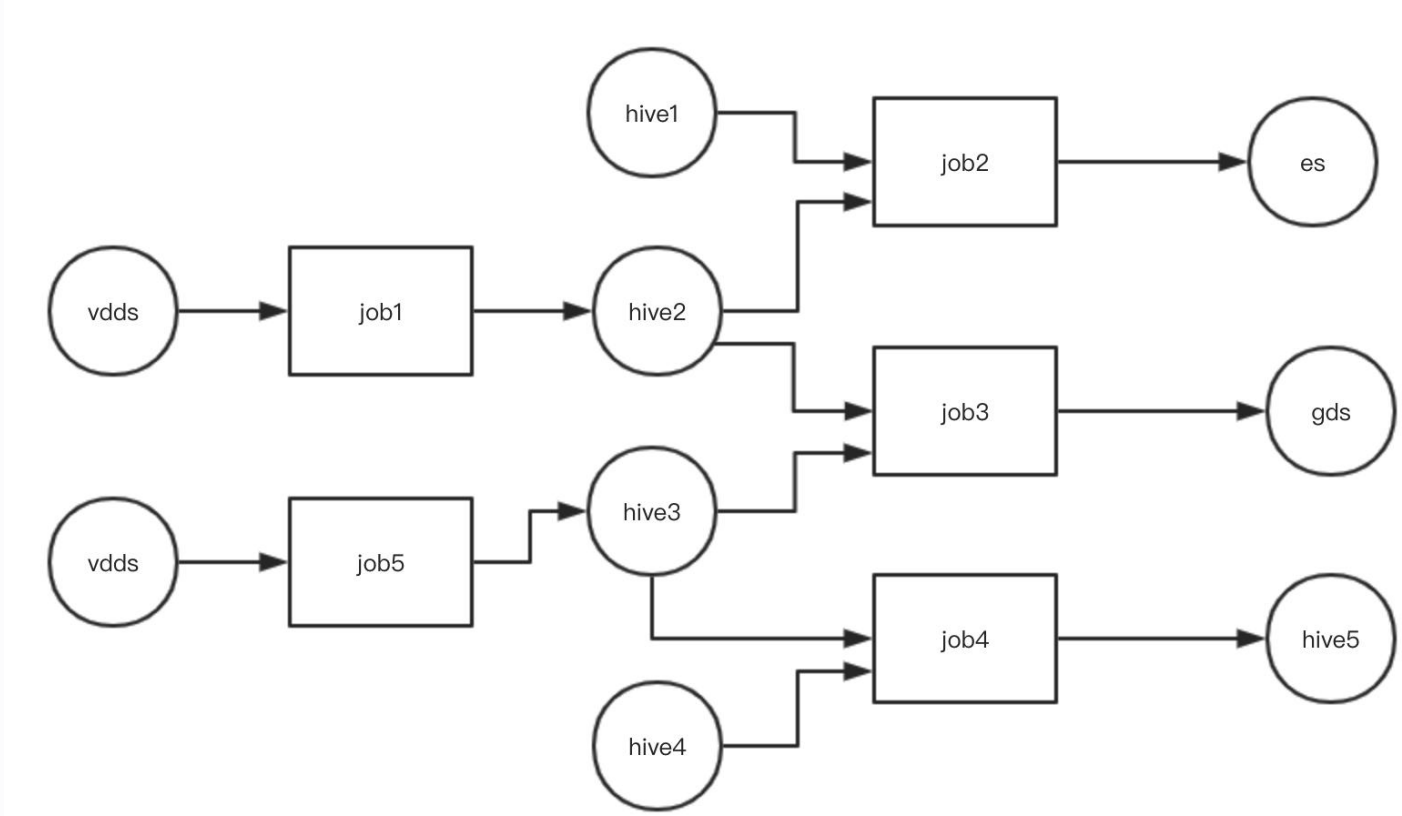
任务的概念和意义

- 数据想要在不同的介质中流转，处理，这个处理数据的东西我们抽象成任务，对于任务来说就是将data[0,n]->data[0,n],从而将数据进行处理。
- puffer、vss、mars、azkaban、vlog、metis等等。
- 任务的唯一标识：任务空间://集群/集群内任务唯一id



数据链路的概念和意义

- 能够清楚的看到每个数据是如何产生的，被哪些任务用了
- 能清楚的知道任务用了哪些数据，生产了哪些数据
- 数据链路的可靠性：**算法+注册数据的全面性和准确性。**

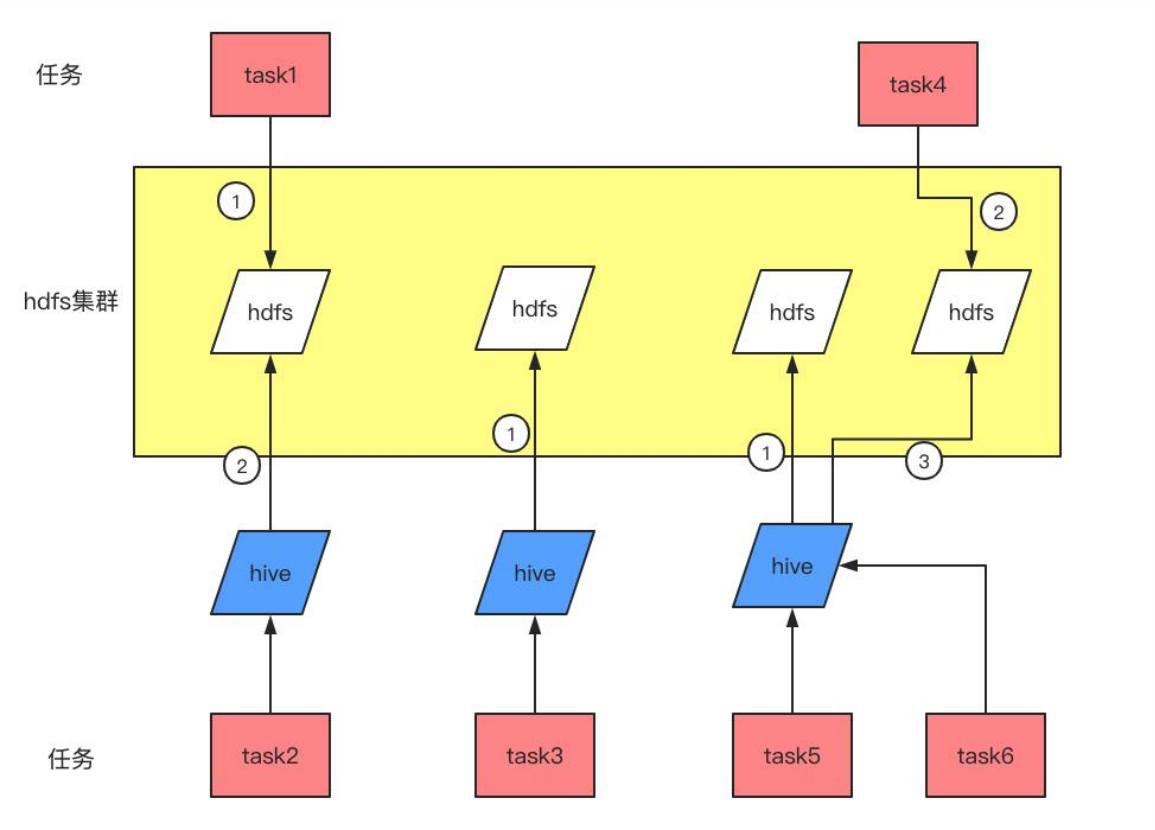


数据全链路在产品上的体现

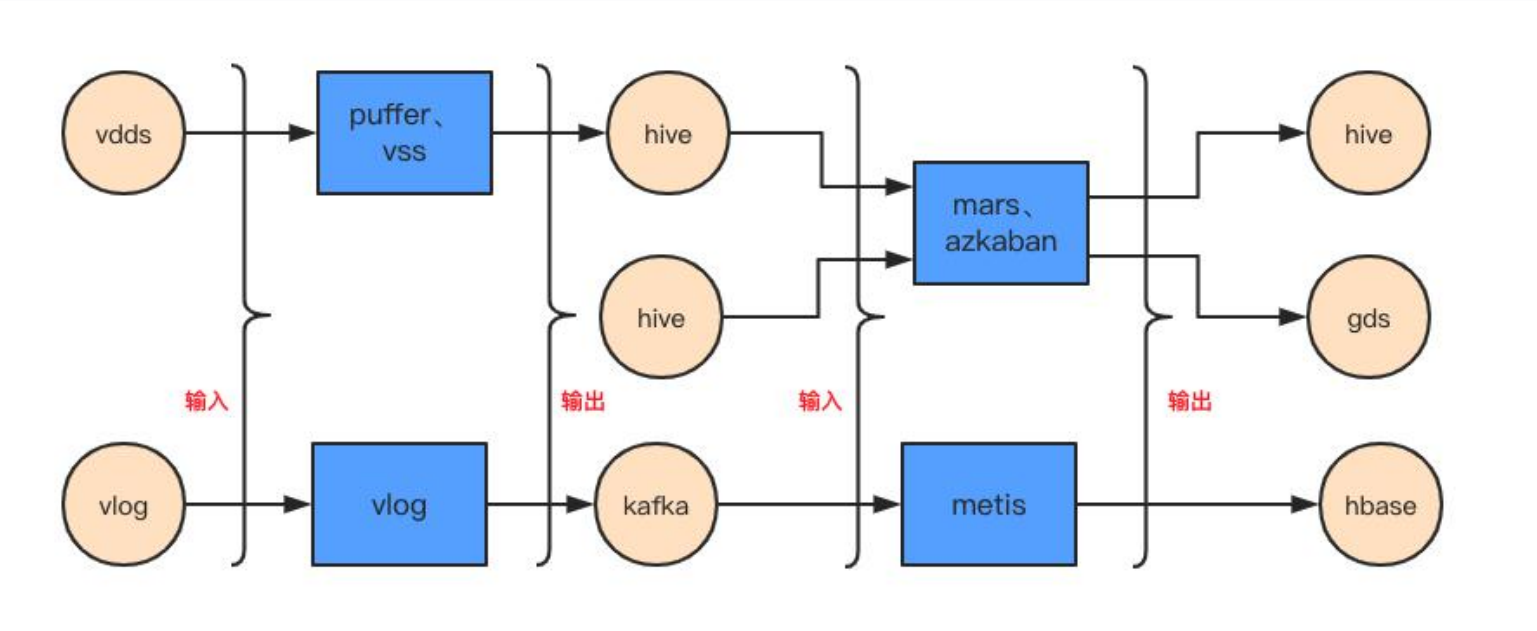
- 各个任务调度系统将自己调度系统的任务处理元数据进行上报。（这个是全链路的基础）
- 查看注册了哪些任务以及任务的**输入输出**是什么？
- 查看涉及到的资源及资源和任务之间的联系。
- **关联资源，ct任务的概念（后面会有多处用到这个概念）**
- 任务血缘、资源血缘
- 任务下线、资源下线
- 产品链接：任务列表入口

关联资源的概念

- 在全链路的设计中，我们创造了关联资源资源和ct任务的概念，使用这二个概念能够将链路的准确性更加严密。
- 关联资源：是指和某个资源息息相关的资源（或者在某种意义上是等同的概念），我们这里主要是指hdfs资源和hive资源处于相互关联的关系。
- hive和hdfs的关系（**相互依赖，多对多**）

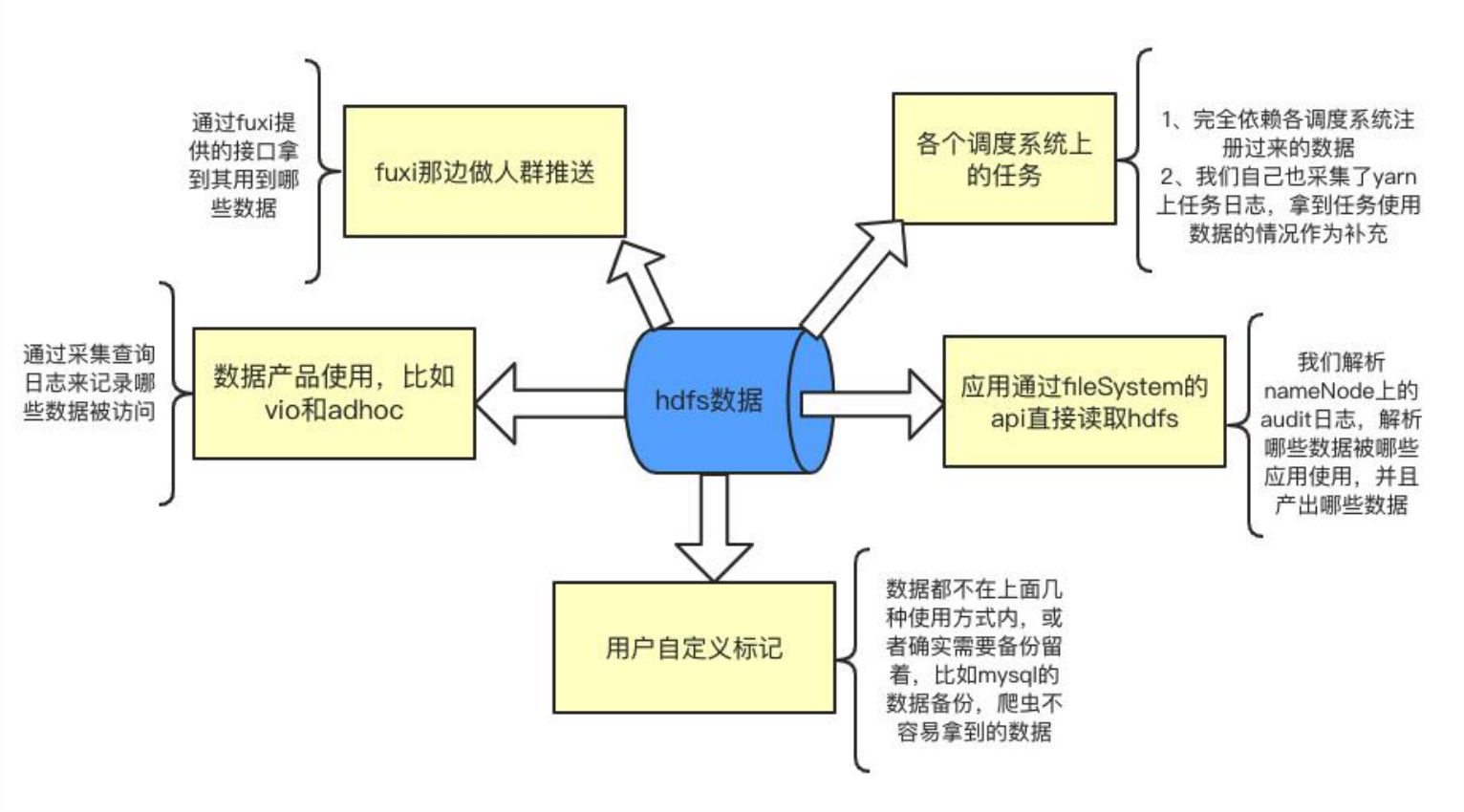


资源和任务之间的联系



- mars任务开发相关手册：
  - mars系统使用说明
  - meta相关的使用文档
  - 接入全链路的文档
  - mars上自定义spark，mr任务配置输入输出

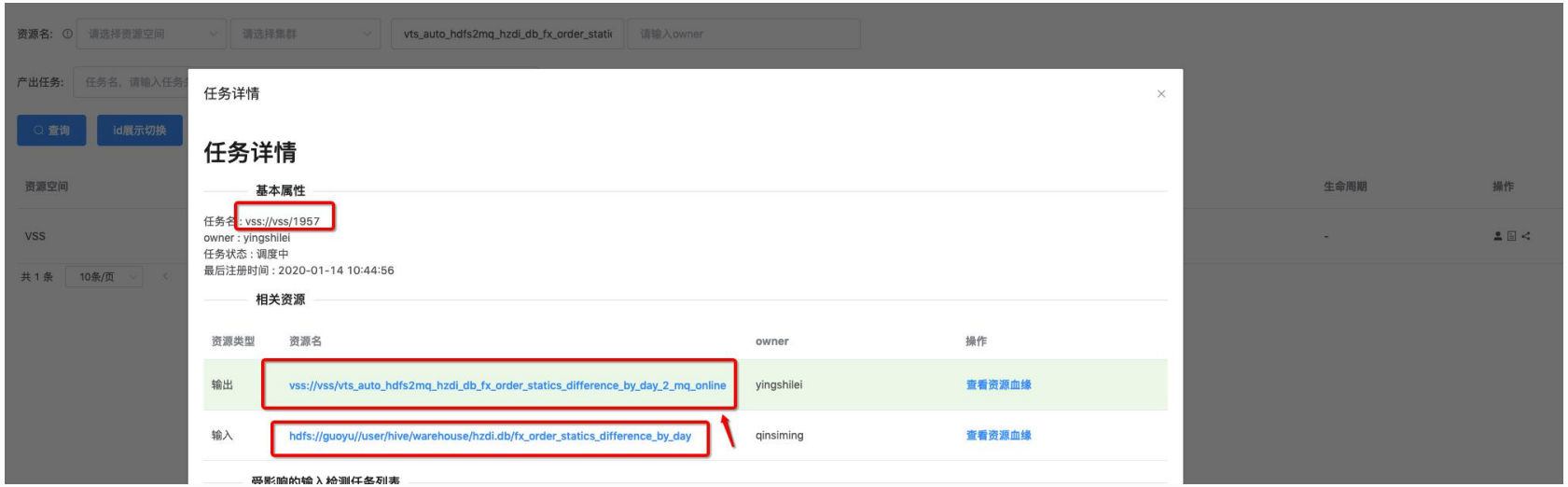
资源的使用方式



- 资源下线和任务下线的逻辑流程
  - 资源下线
  - 任务下线，点击查看

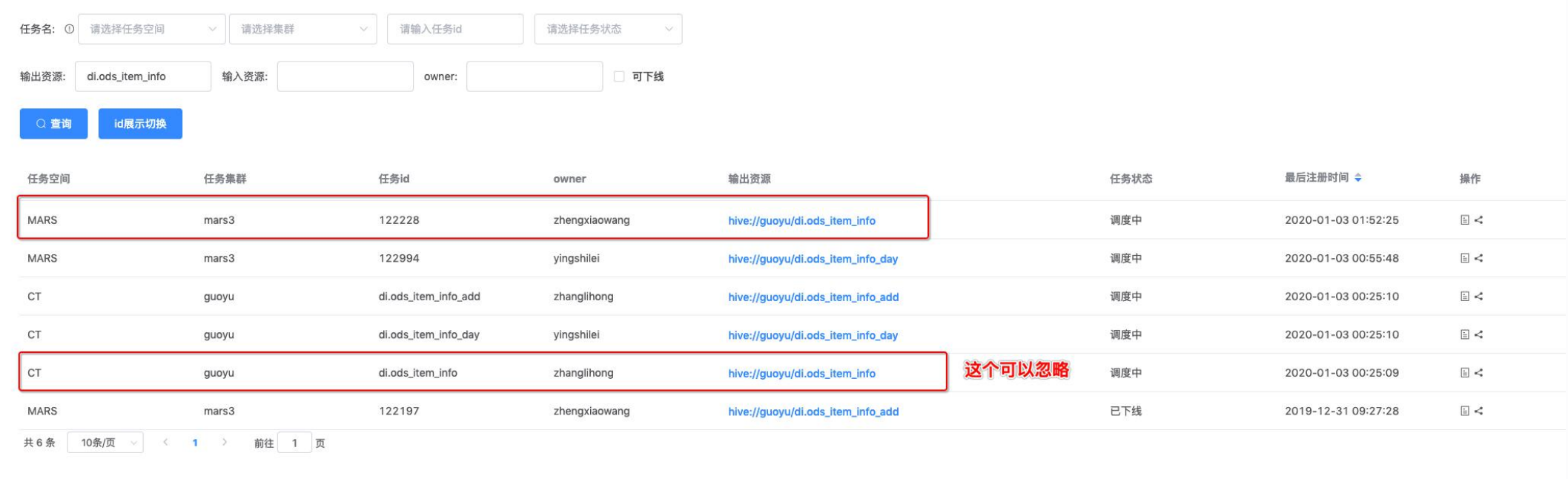
我这个数据更新了，需要更新下游数据，我要重跑哪个任务

- 背景：
  - 有人配置了一个任务，将某个离线表的数据导入到mq里面，结果某天发现离线表的数据出错了，纠正后想重新发送一遍消息，不知道要重跑哪个任务，如何解决？
- 如何使用全链路来快速解决？
  - 首先要找到对应的任务，根据topic去搜输出该topic的任务。
  - 找到该任务之后，看任务是哪个系统的哪个任务，负责人是谁，然后找他重跑一下就好了。
  - 以刘浩之前问的一个来举例子：
    vts_auto_hdfs2mq_hzdi_db_fx_order_statics_difference_by_day_2_mq_online



这个表是谁负责的，哪个任务产出的，最近产出有点问题，该找谁帮我看看啊

- 背景：
  - 我使用了表di.ods_item_info，但是现在好像这个表有点问题，是哪个任务产出的这个表，该找谁看？
- 如何使用全链路解决这个问题？
  - 在输出资源那里输入该表即可，搜出来的结果可以忽略ct任务。

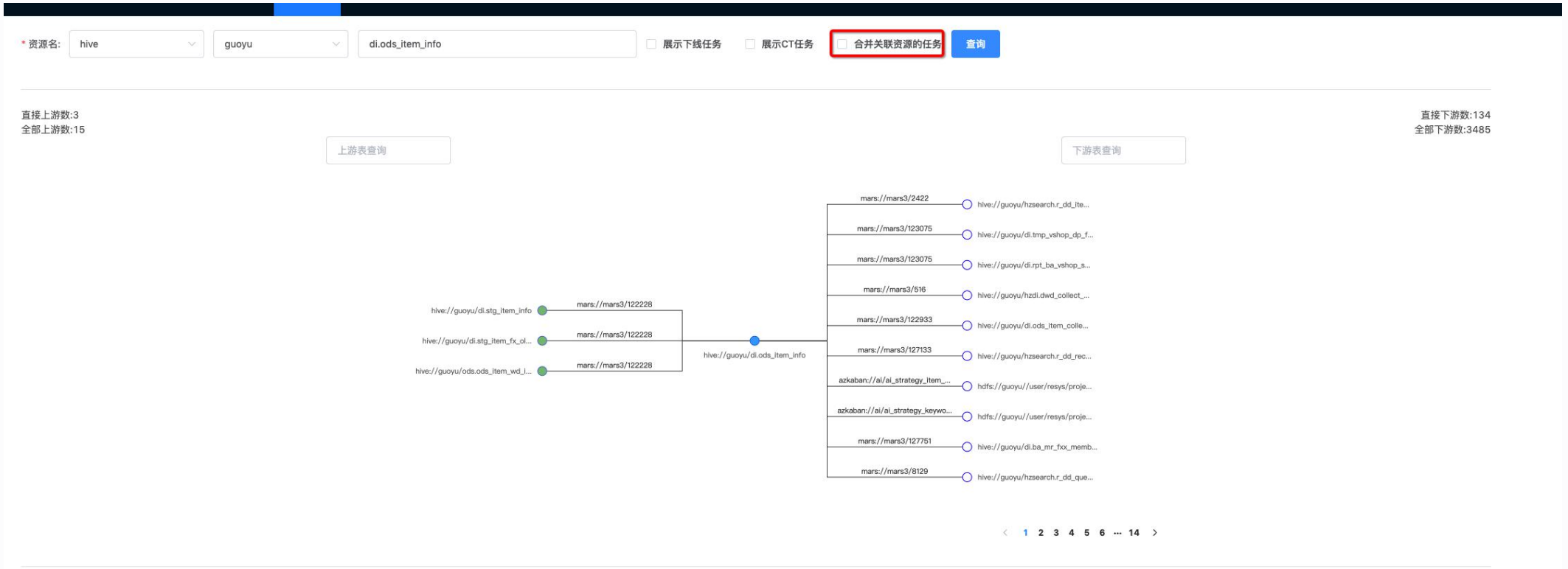


我想知道某个业务库对应的离线表是哪个，怎么看？

- 背景：
  - wd_seller_info.seller_order_info这个表有对应的离线表吗，如果有的话，是哪个？
- 如何使用全链路来快速找到该离线表？
  - 分析：这是一种db->hive的任务来搞定的，所以该业务库肯定是作为输入资源的。
  - 而且是vdds，我们需要在列表页根据做相关的条件即可。

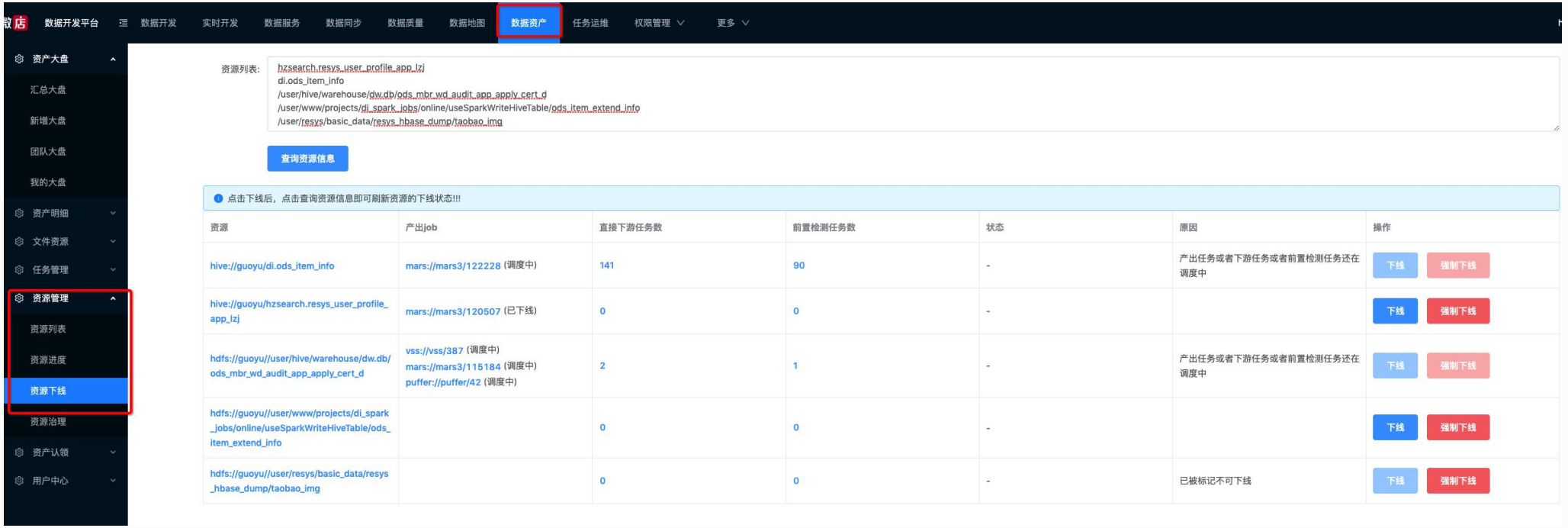
我需要做重构，A表需要下掉，需要使用B表,有哪些下游任务需要切呢

- 背景：
  - 数仓要做重构或者老表到新表的切换，A表不用了，要用B表替换，下游影响多少表，哪些任务需要改代码做切换？
- 如何快速使用全链路解决这个问题？
  - 重点理解：下图中红框标起来的地方。（**合并关联资源的任务**）



我想下掉某个表或者某个hdfs可不可以下掉？

- 背景：
  - 由于某种原因，我有一张表或者hdfs已经无用了，为了避免再被别人下掉，我想把这个资源下掉，但是我该如何判断这个资源确实已经没有被别人用并且顺利下掉呢？
- 如何使用全链路来解决这个问题？
  - 可以在全链路的资源下线页面来查看是否可下线，以及不可下线的原因以及顺利下线。



任务相关的痛点

- **我们团队quota不够用，任务分配不到资源，该怎么弄？**
  - 分配不到资源->团队quota不够用，那就需要下线无用的老任务，也就是我们每周推的下线无用任务，大家需要积极配合，为了给自己团队的新任务有足够的计算资源，初步可以到任务血缘那里看任务有没有下游。
- **我这个任务失败了，影响了多少下游任务，该怎么看？**
  - 这个也是归结到任务血缘那里，看任务的直接和所有下游个数即可。
- **集群管理者，最近集群的计算资源很紧张，需要下批任务，看哪些任务可以下？**
  - 也是归结到无用任务下线的问题。
- **业务下线，DBA想释放db资源，下游相关的大数据任务和线上业务哪些任务需要搞掉？**
  - 也是归结到任务下游的问题。
- **vss任务下线了，为啥半夜还会受到告警电话？**
  - 前置检测任务没有下掉，查看相关路径的直接下游和相关的前置检测任务即可。

后续改进和优化

- 可以实时看到任务可不可下线状态
- 将用户注册的链路信息和我们解析yarn上的任务的输入输出信息做对比，给出不一致列表
- 自动将mars上还没注册的任务的输入输出信息进行补充。
- 提高系统的可用性和易用性。
- 欢迎大家提优化建议。