目的

数仓是商业智能的基础，它为OLAP、数据挖掘提供分析和决策支持。本文以在声波业务中的实践经历，总结了如何开始构建一个数仓模型、如何配置数据任务流调度、以及如何在自助取数上抽象模型配置cube！

声波app是网易云音乐推出的一款主打语音交友的陌生人社交软件，能够进行语音连麦、1V1聊天、娱乐交友等互动的平台。

1、声波基础数仓模型

1.初期设计

早期的声波数仓模型

数据域/业务过程	进房	发言	开播	送礼	提现	充值	上麦	分享	关注	留存
观众	✓	✓		✓	✓	✓	✓	✓	✓	✓
主播	✓	✓	✓	✓	✓	✓	✓	✓	✓	
直播间	✓	✓	✓	✓		✓	✓	✓	✓	
礼物	✓		✓	✓						

早期的声波数仓模型梳理了主要的业务域以及核心业务过程，构建的总线矩阵如上图所示，当然现在看来是有很多不够完善的地方

2.思考

因此，当有机会接手一个新APP时，该如何构建一个完整的数据仓库模型？这里总结了相关经验以及踩过的坑，建议按照以下步骤进行：

step1：数据调研

数据调研步骤非常重要，主要目的是确定需求及需求分析。可以通过调查或访谈等形式来了解，主要分为两部分。

1）咨询不同需求方对数据仓库的需求

不同分工人员（分析师、策划、运营、财务等各个部门）对数据仓库的需求不同，前期需要咨询他们的初期、中期、长期的目标，了解了目标，才能够建设有利的数据仓库。以下是声波业务的需求归纳：

- 分析师的初期的期望是能够产出自上而下的，可直接监控业务大盘的数据，中长期是对新的产品功能做监控，帮助业务实现营收KPI；
- 策划在初期的期望是能够对用户在注册登录环节的流失做些分析，中期是不断扩展拉新用户方式，并对用户做渠道归因，同时增加新的产品功能体验，提升用户留存，长期目标是引导用户付费，实现年终营收KPI，需要分析相关数据；
- 运营的目标是为产品提供运营抓手，组织月度活动，以及对厅主做培训，引导用户付费，其中需要分析活动效果及厅主培训成果；
- 财务的需求是按月要求输出各类型用户的消费、收益、充值、毛利率预估、波币的movement等监控数据。

（2）整体的业务数据框架

这里按照AARRR模型整理了声波业务的整体框架，如下表所示：

生命周期	大类	项目	应用场景
获取(A)	渠道	不同拉新渠道转化转化监控	渠道用户及设备的激活注册转化情况eg： 域内合作、第三方付费、KOL、人拉人、自流量
激活(A)	活跃	活跃用户转化监控	整体活跃用户的后续活跃-观看转化、观看-互动转化、互动-付费转化情况
留存(R)	留存	用户粘性	活跃用户、观看用户、付费用户后续的活跃留存，以及相关功能的功能留存监控
收益(R)	营收	用户营收监控	消费、充值、收益、提现
	流量	曝光、点击、观看转化监控	进房位置的转化监控
	功能	版本迭代新功能、月度活动监控	SayHi功能、在线匹配功能、动态功能
	用户	用户粒度数据宽表	不同类型用户明细，从活跃、生产、消费维度
	房间	房间粒度效率转化	监控房间维度效率转化

step2：主题域

主题域一般可以按照企业的部门划分，也可以按照业务过程或者业务板块中的功能模块进行划分，这里遵循云音乐主端的规范，将一级主题域划分为参与者、服务及产品，版权及协议、公共、事实这5个大的主题域，二级细节分类在下文中详述。

step3：定义维度与构建总线矩阵

维度建模中我们选用了Kimball维度建模方法，在定义好主题域之后，需要对具体对业务过程做分类。

下面是对声波重新构建了总线矩阵，结构大致如表所示：

			一致性维度							
一级数据域	二级数据域	业务过程	用户	房主	房间	礼物	设备	动态	坑位	进房入口
参与者	用户		✓							
参与者	房主			✓						
参与者	房间				✓					
事实	交易营收	消费	✓	✓	✓	✓	✓			
事实	交易营收	收益	✓	✓	✓	✓	✓			
事实	交易营收	充值	✓	✓	✓	✓	✓			
事实	交易营收	提现	✓	✓	✓	✓	✓			
事实	社交互动	进房	✓	✓	✓		✓−		✓	✓
事实	社交互动	出房	✓	✓	✓		✓−		✓	✓
事实	社交互动	发言	✓	✓	✓		✓−			
事实	社交互动	上麦	✓	✓	✓		✓−			
事实	社交互动	下麦	✓	✓	✓		✓−			
事实	社交互动	分享	✓	✓	✓		✓−			
事实	社交互动	关注	✓	✓	✓		✓−			
事实	社交互动	收藏	✓	✓	✓		✓−			
事实	社交互动	匹配	✓	✓	✓		✓−			
事实	社交互动	私信	✓	✓	✓		✓−	✓		
事实	社交互动	点赞	✓				✓−	✓		
事实	社交互动	评论	✓				✓−	✓		
事实	日志流量	曝光	✓	✓−	✓−		✓		✓	✓
事实	日志流量	点击	✓	✓−	✓−		✓		✓	✓
事实	日志流量	浏览	✓	✓−	✓−		✓		✓	
事实	日志流量	启动	✓	✓−	✓−		✓		✓	
事实	营销活动	push	✓							
事实	营销活动	拉新	✓							
事实	安全风控									
事实	搜索									
事实	AB测试									
服务及产品	UGC							✓		
服务及产品	PGC					✓				
服务及产品	公共资源								✓	✓
服务及产品	特权服务									
服务及产品	音乐协议									
服务及产品	直播协议									
公共	日期									
公共	设备						✓			

其中打勾的是指在对应的业务过程功能与一致性维度下有关联性，由此来构建事实表与维度表。

step4：明确指标统计

统计指标一般是来源于分析师梳理的监控报表。其中指标=时间周期+统计粒度+业务过程的度量（描述）。请注意，口径一般包含业务口径和技术口径。一般业务口径分析师经与策划等沟通后会给到一个明确统一的口径，技术口径可能需要我们回填给业务。

声波的常用统计指标一般包括（时间周期）近1/3/7/15/30日的（统计粒度）用户的（业务过程）进房次数、进房时长、消费金额、是否留存等等。

step5：结果验证

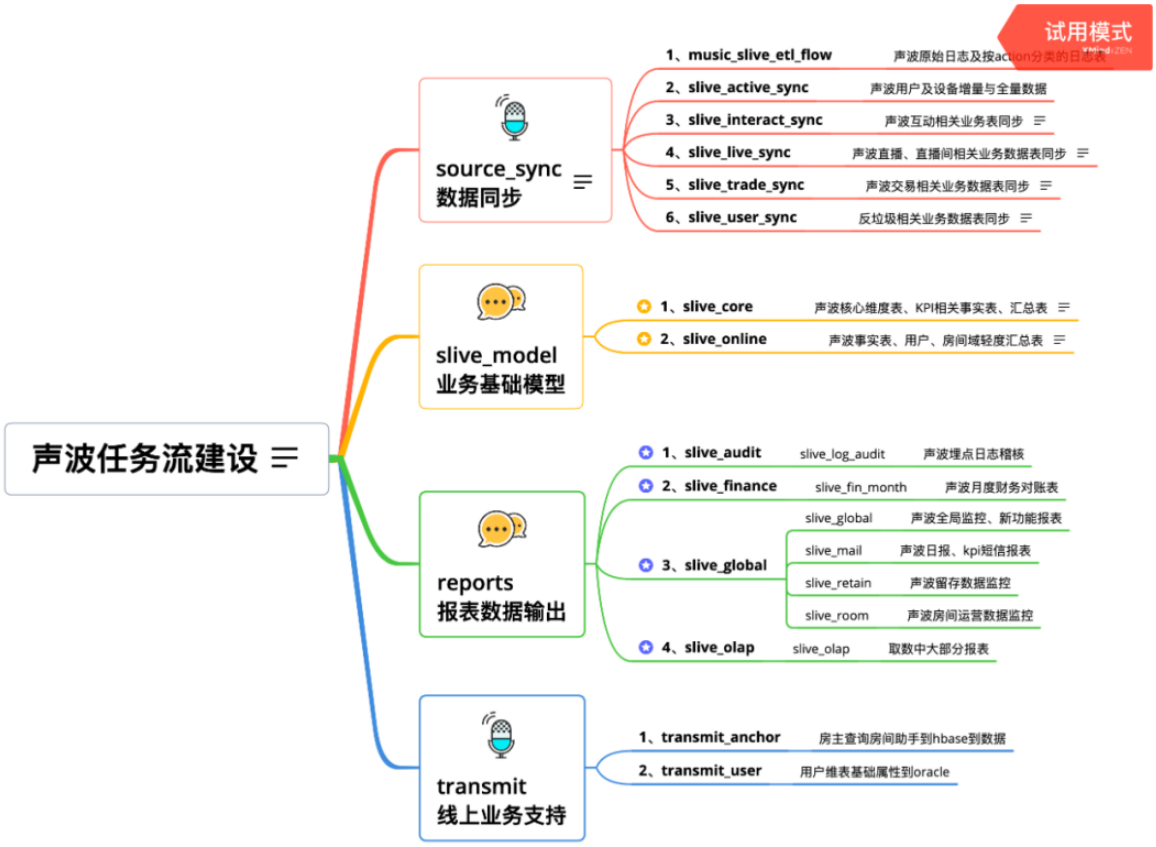
构建数据仓库，主要是为我们及下游使用方分析数据，以赋能产品，所以数据的准确性至关重要，可以先通过网易有数大数据平台提供的数据测试中心的测试功能进行测试、找前辈们review代码或者通过与已有报表中相同指标进行核对，然后再交由分析师配置报表。

至此，我们将数据做了一个以业务过程为分类的纵向划分，下面要开始对数据做横向的分层。

2、声波任务流建设

2.1.现有设计

声波数据任务流的地址为：XXXX。主要分为source_sync、slive_model、reports、transmit四个目录。下图中详细描述了各文件存放的数据内容



2.2思考

如何进行数据分层?

常规的数据仓库模型分层一般包括ODS贴源数据层，DIM维度层，DWD明细数据层，DWS轻(中/重)度汇总层以及ADS应用数据层。在分层中我们需要思考以下问题：

业务数据是根据什么（维度、粒度）汇总的，衡量标准是什么?

eg:声波划分的粒度包括：用户粒度、房间粒度、用户+房间粒度、动态粒度、用户+动态粒度，衡量标准主要是各粒度下的各种指标：次数、时长、金额等等。

DIM该如何设计？DWD和DWS应该如何设计？是否有公共的指标?

eg:DIM是观察业务的角度，建议可以设计主维表（一般为直接从业务库中同步来的、包含常用的属性、稳定性高）和多个次维表（常变更），因为维表一般有严格的时间要求与依赖任务，同时对与需要经常修改的任务做回跑有利，例如我们为声波用户设计了基础的dim_slive_user_base_d主维表和指标丰富的dim_slive_user_d次维表。

DWD一般是基于具体业务过程，构建最细粒度的明细层事实表，也可以将明细事实表的某些重要维度属性字段做退化设计，这里也可以添加一些常用统计口径的杂项维度。例如消费明细表中并不是每一笔消费都是计入KPI的，可以添加一些is_real_consume（是否实际消费）、is_consume_water（是否消费流水）、is_operation（是否运营操作），这样便于下游的统计。

DWS以分析的主题对象作为建模驱动，将相对应的事实进行聚合统计，形成一些轻度聚合、中度聚合或者重度聚合的宽表。

任务调度设置

任务流建设是对整个业务过程的横向分层，下表是声波业务的任务分层，也是契合于上述的数据分层。

分层	任务流	描述	依赖配置	调度时间	执行设置
ods	source_sync文件夹所有任务流节点	日志同步、业务表同步	DS日志归档、pandora业务表	1点到3点	开启高优先级、自动重试、失败设置为继续完成其他分支节点、并发执行
dim/dwd/dws	slive_core	KPI相关、大盘监控报表相关	依赖source_sync中所有节点	3点	配置电话失败、超时报警、开启高优先级、自动重试、失败设置为继续完成其他分支节点、并发执行
dim/dwd/dws	slive_online	新功能的中间层表建设	依赖slive_core	3点30	开启高优先级、自动重试、失败设置为继续完成其他分支节点、并发执行
ads	reports文件夹下slive_mail	KPI/大盘	依赖slive_core	4点45	配置电话失败、超时报警、开启高优先级、自动重试、失败设置为继续完成其他分支节点、并发执行
ads	reports文件价下除slive_mail外任务流节点	有数报表	依赖slive_online	4点到5点	开启高优先级、自动重试、失败设置为继续完成其他分支节点、并发执行

优缺点：

- 优点：相对于单表单任务流来讲，任务流依赖配置简单，同时易读性高；对于有依赖的任务可一次性提交回跑任务；核心任务能够及时产出。
- 缺点：相对于单表单单任务流来讲，定位某表属于那个任务流相对较慢；某些核心任务报表产出相对延迟些。

注意：

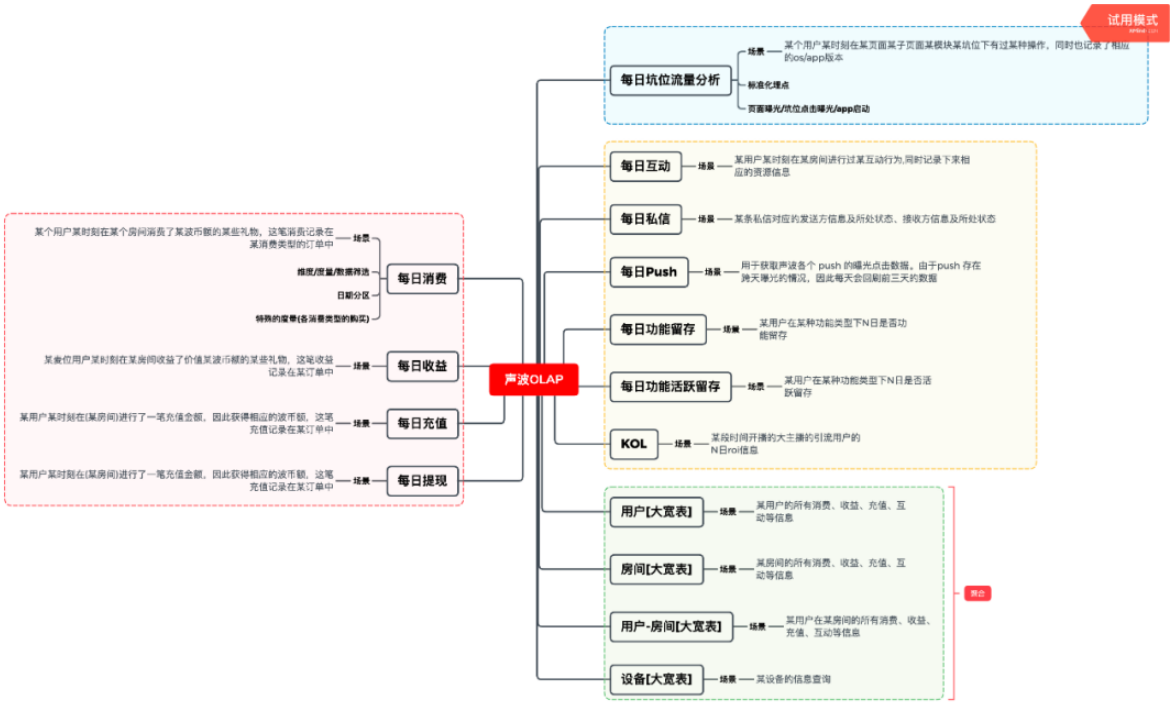
- 当调度任务失败时，要在实例详情页面进行重跑；
- 一般依赖任务的开始调度时间要晚于被调度的任务；
- ads层表从dim/dwd/dws层来调用，不要直接走ods层。

3、声波OLAP取数模型及流量自动化

3.1现有设计

快速便捷获取高质量数据是业务侧的希冀，同时为减少ad-hoc式的查询也是我们的希冀。因此构建声波olap模型，目标是帮助业务人员快速使用数据，获取结果并用于业务生产。

取数模型设计：



目前模型建设流程：

1. 业务侧/已有报表中归纳常用指标
2. 模型设计（常分析的业务过程（交易、互动）的明细、用户、房间、用户+房间的汇总表）
3. 模型评审
4. 配置模型（使用场景的说明、字段业务口径、技术口径、添加自定义维度、自定义度量）
5. 测试使用

流量自动化的解决方案：

1. 策划梳理坑位信息-〉与策划勾兑坑位信息-〉设计埋点scm信息-〉上传埋点到埋点平台-〉与开发勾兑埋点内容--〉下载最终版坑位信息制作坑位码表。
2. 设计流量自动化的聚合表模型：uid+os+appver+mspm+source（+房间id+房间模版类型id）的粒度统计对应的(曝光、点击、进房)次数、人数和时长等信息。构建流量自动化模型，最终可由取数

展示，该模型可以查看日常曝光点击的坑位PV、UV，同时可以查看核心（曝光-点击-进房）漏斗数据。

3.2思考

目前存在的难点，这些都是后期会优化的内容。

- 数据侧：a.模型设计(聚合、解耦) b.模型迭代回跑
- 平台侧：a.平台开发进度无法满足模型使用 b.问题响应速度依赖其他部门
- 业务侧：a.对数据的维度、度量、聚合、日期分区的理解困难 b.自主分析数据的习惯尚未建立