基本概念

元数据

元数据是描述其它数据的数据，主要描述数据的基本信息，例如数据的存储位置、大小、存储方式、创建时间等。

数据标准

数据标准是指保障数据的内外部使用和交换的一致性和准确性的规范性约束。

数据元

由一个属性集合规定其定义、标识、表示和允许值的一个数据单元，作为数据标准的实际载体。

对象类

现实世界中的想法、抽象概念或事物的集合，有清楚的边界和含义，并且特性和其行为遵循同样的规则而能够加以标识。

特性

对象类的所有个体所共有的某种性质。

语境

一个名称所用于的或所源自的应用环境或规程的描述。

值域

允许值的集合。

数据字典

表示某一数据项的取值枚举集合，在规定的集合里取值，一般由代码值，代码描述组成一个字典项。

原始数据字典

原始数据字典用于表示原始数据（来源数据）的某一数据项的枚举集合。

标准数据字典

标准数据字典用于表示经由标准规范约束的数据的某一数据项的枚举集合，一般作为数据元的值域表示。

角色&成员

成员是具备访问或使用数据开发及管理平台的账号，添加成员前需要在账号管理系统中事先注册，添加成员时可对成员赋予角色。

角色是项目内一系列权限的集合，把角色赋予成员后，成员即具备了角色所有权限。成员可以同时拥有多种角色。

项目中内置的角色包括：负责人、管理员、指标管理者、数据团队管理者、指标审批者。

修饰词&衍生词

修饰词是对指标进行限定抽象的业务限定，修饰词归属于一种修饰词类型，比如日志域的访问终端类型，包含修饰词PC端、无线端等。

衍生词用于修饰原子指标，是对于原子指标中带有计算口径的词进行了抽象定义。

原子指标

原子指标是有业务统计含义的数值型数值，通过度量加工得到。

派生指标

派生指标 = 原子指标 + 修饰词 + 时间周期

衍生原子指标

衍生原子指标 = 主原子指标 + 衍生词

复合指标

复合指标由一个或多个派生指标通过计算而成。

业务口径

从业务的角度制定统一的数据统计标准，往往用来说明某一数值在特定业务场景下的含义，例如新增用户数、活跃用户数。

技术口径

用来描述某一数值（字段）通过其它字段加工得到的计算逻辑。

指标域

指标域属于指标管理的上层分类概念，可以把比较接近的业务过程或者属性相近的内容划分为一个大的整体，称之为指标域。

流水型

数据传输中，一种数据过滤的方式，主要应用在无法通过日期字段进行增量同步的场景。

流水型：从选择字段的起始值开始读取数据，读取到最新记录位置，下次从上次的最新记录读取至当前的最新记录。

依赖节点

节点A计算运行时需要使用节点B的计算结果，则认为节点A依赖于节点B，节点B即为节点A的依赖节点。

跨周期自依赖

节点计算当前任务时必须依赖此任务的上一周期实例。

主题域

数仓建设的一个上层分类概念，把比较接近的业务过程或者属性接近内容划分为一个大的整体，称之为主题域。

维度

维度是度量的环境，用来反映业务的一类属性，常见的如统计日期、用户、省份、性别等。

度量

来源于业务系统中不经过加工的用于反映和描述事实的数值型数据，不带业务口径。

维度表

维度表包含了事实表中指定属性的相关详细信息。

桥接层

用于存放桥接维度的关系表。

明细层

从ods层经过ETL得到的明细数据，表示具体的事实，主要由维度和度量等构成。

汇总层

由明细数据经过汇总得到的数据，主要由统计维度和指标构成。

应用层

由明细层或汇总层加工得到用来面向报表、服务、应用等使用的数据。

贴源层

由业务系统同步到数据仓库的原始数据，一般不经过加工。

实例

任务一般需要生成实例后才会运行，任务和实例的关系是一对多，任务包括静态代码和配置，只有生成实例才会将代码逻辑真正运行起来。

重跑

任务可多次运行，每次运行可称为重跑，一般在任务异常、数据有误、逻辑有变更时需要重跑任务。

补数据

当任务异常或逻辑有变更时，可对任务选择历史一段时间来进行补数据，以便修正历史数据。

数据血缘

属于元数据的一部分，用来展示数据表之间的链路关系，包含了数据的来源、加工方式、映射关系以及数据去向。

活跃血缘

对于离线开发场景，活跃血缘指的是线上调度产出的血缘，且调度持续生效；对于实时开发的场景，活跃血缘指的是数据由实时计算平台定时推送，推送时刻如果为运行状态。

静默血缘

对于离线开发场景，静默血缘指的是未来可能不会更新或即将失效的血缘，包括开发模式运行、线上调度已运行过但是已取消调度、线上模式严重逾期执行、自助分析运行SQL等；对于实时开发场景，静默血缘指的是数据由实时计算平台定时推送，推送时刻如果为未运行状态。

基线

即时间线，在基线运维场景中，通过设定几条时间线（即基线），并将线上任务关联到这些时间线上，当这些任务的实例或上游实例运行失败，或者这些实例的预计产出时间超过对应的时间线时，系统会触发报警通知给值班人员。

预警

指预警时间，表示当前基线配置的预警时间线。当基线的任务实例预计完成时间超过“预警时间”，则会发送报警类型为“基线报警”的报警。

破线

指破线时间，表示当前基线配置的破线时间线。当基线的任务实例预计完成时间超过“破线时间”，则会发送报警类型为“基线报警”的报警。

形态探查

对Hive表进行数据形态的探查，一般用于新接入的源头表或者新加工出来的数据的初步探查，了解数据整体的情况。也可以通过探查，发现数据的潜在问题，比如主键唯一性、字段空值、非常规字段值等等。

数据比对

对两个hive表进行逐字段级别的比对，一般用于重要表的加工逻辑调整后或者模型重构后，比对原有数据，确保加工出来的新数据和原有数据的一致性。如果是加工逻辑调整，比对结果一致或差异可解释并能接受，则可认为新的逻辑可以提交上线；如果是模型重构，则认为新的模型数据正确，可安排老模型下游迁移等。