ICS 03.060
CSS A 11

JR

中 华 人 民 共 和 国 金 融 行 业 标 准

JR/T XXXXX—XXXX

投资研究时序数据参考模型
Standard investment analysis referencing data model

（征求意见稿）

XXXX－XX－XX 发布　　XXXX－XX－XX 实施

中国证券监督管理委员会　　发 布

目 次

前言 1
引言 2
1. 范围 3
2. 规范性引用文件 3
3. 术语和定义 3
4. 投资研究主数据逻辑模型的数据分类 4
4.1 主数据定义与描述信息 4
4.2 主数据实体信息 5
4.3 主数据参考信息 5
4.4 投资研究主数据表清单 5
5. 投资研究主数据逻辑模型图 6
6. 主数据模型数据表详细设计 6
6.1 MDB_DEFINE 定义表 6
6.2 MDB_DEFINE_ATTR 主数据属性表 7
6.3 MDB_DEFINE_RELA 关系定义表 7
6.4 MDB_ENTITY_INST 机构表 8
6.5 MDB_ENTITY_SECURITY 证券表 10
6.6 MDB_DEFINE 定义表 10
6.7 MDB_ENTITY_REGION 地域表 11
6.8 MDB_ENTITY_PERSON 人物表 12
6.9 MDB_ENTITY_CONCEPT 市场概念表 12
6.10 MDB_ENTITY_PROD 产品表 13
6.11 MDB_ENTITY_PROD_CLS 产品分类表 13
6.12 MDB_ENTITY_PROD_CLS_ATTR 品类属性表 14
6.13 MDB_ENTITY_PROD_PROP 产品属性值表 14
6.14 MDB_ENTITY_DICT 主数据实体别名表 15
6.15 MDB_ENTITY_COMMON 通用实体表 15
6.16 MDB_ENTITY_COMMON 通用实体表 15
7. 投资研究主数据逻辑模型应用场景 16
7.1 场景一 数据表意消歧 16
7.2 场景二 指标表达标准化 16
7.3 场景三 数据智能应用 17
8. 指标标准模型设计方案 18
8.1 背景 18
8.2 标准化模型设计 19

8.2.1 UIDS_IND_DEF（原子指标定义表） 19
8.2.2 UIDS_IND_DIM（原子指标维度表） 20
8.2.3 UIDS_IND_DER_DEF（派生指标定义表） 21
8.2.4 UIDS_IND_DER_VALUE（派生指标数值表） 22
8.3 派生指标命名规范 22
9. 维度表达标准化设计方案 23
9.1 维度类型编码（主数据）规范 23
9.2 通用维度码值规范 24
9.2.1 证券及相关金融工具 security 24
9.2.2 公司 organization 25
9.2.3 地域 region 26
9.2.4 人物 person 26
9.2.5 产品 product 26
9.2.6 行业 industry 27
9.2.7 渠道 channel 27
9.2.8 币种 currency 28
9.2.9 统计方式 calc_prop 28
10. 通用指标模型 ETL 作业的设计方法 28
10.1 概述 29
10.2 方法简述 29
10.3 通用指标配置模型详述 30
10.3.1 UIDS_DATASET_DEF（源数据集定义表） 30
10.3.2 UIDS_IND_DEF（原子指标定义表） 31
10.3.3 UIDS_IND_CONFIG（原子指标配置表） 31
10.3.4 UIDS_DATASET_DIM（源数据集维度定义表） 31
10.4 ETL 作业步骤简述 32
10.5 模型验证 33
参 考 文 献 34

前　言

本文件按照GB/T　1.1—2020《标准化工作导则　第1部分：标准化文件的结构和起草规则》的规定起草。

本文件由全国金融标准化技术委员会证券分技术委员会（SAC/TC　180/SC4）提出。

本文件由全国金融标准化技术委员会（SAC/TC　180）归口。

本文件起草单位：中国证券监督管理委员会科技监管局、嘉实基金管理有限公司、中证信息技术服务有限责任公司、银华基金管理有限公司、中国国际金融股份有限公司、中信证券股份有限公司、中国人寿资产管理有限公司、资本市场学院、上证所信息网络有限公司、基金业协会。

本文件主要起草人：姚前、蒋东兴、刘志明、周云晖、路一、刘彬、李大炎、刘瀚月、梅亚雷、徐宇、张淼、黄建、徐倩、彭乔、陈思遥、杨琳、李珊珊、张若海、宋广超、罗俊、高贵中、刘雪峰。

引　言

近年来，随着信息科技与传统金融业的结合日趋紧密，越来越多的数据深度参与金融业日常业务流程已是大势所趋。投资研究是金融领域的重要环节，其涉及的数据范围之广，专业程度之深，数据非标准化程度之高已成为数据的治理和流通的绊脚石。

目前金融行业内，在证监会的指引下，传统的证券期业务已开始了数据模型标准化的工作[1]。对于投资研究领域，其数据化程度相对较低，数据形态多样，机构内及机构间数据交换频繁、业务发展迅速。目前因行业内缺乏数据流通标准，各机构、数据公司均按照自己的标准生产和分发数据，造成数据使用者往往需要重复建设数据采集和标准化程序。另一方面对于数据生产者来说，往往需要不同结构的数据结构需求重复加工数据。如此造成了 N×N 的复杂、低效的投研数据生产和传输模式。

为提高数据交换效率、规范行业机构数据应用系统建设、提升行业数据标准化水平，嘉实基金等资管、券商机构联合组织开展了投资研究数据模型建设工作，旨在清晰描述整个经济生产制造领域的数据流向、数据名称、数据定义、结构类型、代码取值和关联关系等，为行业机构内部系统建设和机构间数据流通提供指导。

本文主要专注于投资研究中使用频率最高的一类数据：时间序列指标的模型。基于此方法能够形成一套容易生产、便流通的模型框架。相关成果是行业数据模型的重要组成，是行业标准的数据审核依据，是行业逻辑模型的映射基础，对于规范行业数据语言、推进行业数据治理、辅助行业监管科技建设等都具有十分重要的意义。

投资研究时序数据参考模型

1. 范围

本文规定了金融投资业中，投资研究领域内的指标抽象模型设计方法，包括总体设计架构、实体标准化方法、通用生产作业设计方法及元语定义。

本文适用于投资机构在投资研究场景中的指标数据抽象模型建设工作。

2. 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 4754—2017 国民经济行业分类

3. 术语和定义

下列术语和定义适用于本文件。

3.1

主数据 master data

满足业务需要的、反映核心业务实体对象状态属性的基础信息。

3.2

逻辑模型 logic model

数据与数据表的逻辑结构，用于描述数据之间的关系。

3.3

指标 Indicator

指标是投资研究中业务人员最常用的数据形态。一组定义清晰、准确的、更新及时的时间序列指标，可以为投资分析人员提供有益的分析、决策辅助。

- 原子指标：指在特定经济环境中、金融实体在某一事件或事实下的度量，具有不可拆分性，具有唯一明确的名称，如产品销售量、采购价、销售价等；
- 派生指标：是指原子指标在不同约束条件（即维度，如时间、地域、公司等）依据数据可得性组合而来的细粒度指标。派生指标通常是业务分析时最常使用的粒度的指标，用于定量地描述某一个实体在特定环境下的客观事实；
- 衍生计算指标：是指在派生指标或原子指标基础上，按照特定的业务分析需求统计计算的指标，例如均值、同比、同比差值等。其中，衍生指标可细分为如下两个子类：
  - ✧ 通用衍生指标：指在派生指标原值的基础上，根据一般性统计方法（例如同比、环比等）获取的加工数据。例如：北京:GDP:同比值；

- ✧ 复合衍生指标：指基于若干基础派生指标，根据行业分析逻辑，投研人员自定义的复合加工数据。例如：全国激光电视市占率 = 全国激光电视销量（月）/ 全国电视总销量（月）。

3.4

实体 entity

实体是参与金融和经济生产活动的对象。因投资业务需要涉及社会经济生活中的所有行业，因此具体到每个细分行业中，有很多具体的垂直行业专属对象，例如互联网行业关注的对象有“APP”、“大 V”，汽车行业关注的对象有“OEM”、“车型车系”，半导体行业关注“制造工艺”等。但是有一部分实体，无论进行宏观、行业还是微观研究，都会经常涉及分析和讨论，本文称其为“通用实体”。经归纳，常见的投研业务的通用实体类型有公司、地域、产品、证券、人物等。这些实体参与经济生产活动的抽象模型如下：

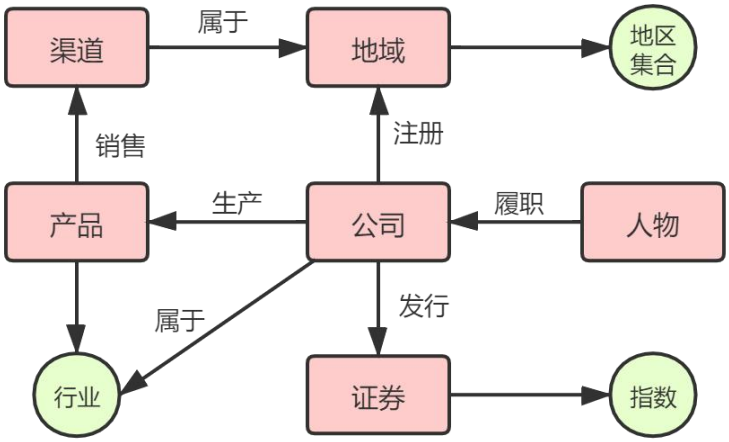


图 1 通用投资研究实体关系示意图

实体可再分为基础实体和组合实体。

- ◆ 基础实体：特定实体类型下不可再分的最小实体单元。例如地域实体类型中的“北京市”；
- ◆ 组合实体：是由若干相同实体类型的实体组成的实体集合。例如地域实体“欧盟”，由 27 个成员国地域实体组成。

3.5

维度 Dimension

维度指描述度量的对象，用来表示在原子指标某一方面的属性。在金融证券行业常见维度即为各种实体或实体的属性，如证券、地域、公司、产品、人等，或产品的口径、产品的规格、产品的销售渠道等。维度还可进一步拆分为标识性维度和非标识性维度。

- ◆ 标识性维度：即数据库语言中的“业务主键”，表示能够决定度量唯一取值的最小维度集合；
- ◆ 非标识性维度：不能决定度量唯一取值的维度，用以描述标识性维度。

4. 投资研究主数据逻辑模型的数据分类

投资研究主数据逻辑模型的数据分类包括主数据定义与描述信息、主数据实体信息以及主数据参考信息三种类别，下述第5.1-5.3节分别为各类别的详细介绍。

4.1 主数据定义与描述信息

主数据定义与描述信息是对投资研究主数据的类型进行定义和描述，同时对每一类型的主数据的属性、不同主数据之间可能的关系进行定义和描述。该类别与主数据实体信息、主数据参考信息之间都存在着密切的关联关系，是逻辑模型的核心内容。

主数据定义与描述信息包括定义、主数据属性、关系定义、关系、产品分类属性、产品属性值以及主数据别名七种类型的数据表（详细内容见第6章）。

4.2 主数据实体信息

主数据实体信息是对投资研究主数据的具体实体进行枚举、编码、定义与描述。例如地域主数据对应的地域主数据实体表，详细列示了所有国家、省州、市、区县等地域行政单位，并进行编码和层级关系描述。

主数据实体信息包括机构、证券、行业、地域、人物、市场概念、产品、产品分类以及通用实体九种类型的数据表（详细内容见第6章）。

4.3 主数据参考信息

主数据参考信息用在描述主数据实体的信息当中。对于范围与取值稳定且需要编码管理的非研究对象类信息，归纳如参考信息管理。例如交易所信息、主数据实体层级信息、机构类型信息等。

主数据参考信息包括数据源参考、机构类型参考、实体层级参考以及交易所参考四种类型。

4.4 投资研究主数据表清单

通过对投资研究主数据逻辑模型的数据分类汇总归纳，形成的投资研究主数据表清单如表1所示。

表 1 投资研究主数据表清单

序号	类别	表名	表释义
1	主数据定义与描述信息	MDB_DEFINE	定义表
2	主数据定义与描述信息	MDB_DEFINE_ATTR	主数据属性表
3	主数据定义与描述信息	MDB_DEFINE_RELA	关系定义表
4	主数据定义与描述信息	MDB_ENTITY_RELA	关系表
5	主数据定义与描述信息	MDB_ENTITY_PROD_CLS_ATTR	产品分类属性表
6	主数据定义与描述信息	MDB_ENTITY_PROD_PROP	产品属性值表
7	主数据定义与描述信息	MDB_ENTITY_DICT	主数据别名表
8	主数据实体信息	MDB_ENTITY_INST	机构主表
9	主数据实体信息	MDB_ENTITY_SECURITY	证券主表
10	主数据实体信息	MDB_ENTITY_INDUSTRY	行业主表
11	主数据实体信息	MDB_ENTITY_REGION	地域主表
12	主数据实体信息	MDB_ENTITY_PERSON	人物主表
13	主数据实体信息	MDB_ENTITY_CONCEPT	市场概念主表
14	主数据实体信息	MDB_ENTITY_PROD	产品主表
15	主数据实体信息	MDB_ENTITY_PROD_CLS	产品分类主表
16	主数据实体信息	MDB_ENTITY_COMMON	通用实体主表
17	主数据参考信息	REF_STANDARD_SRC	数据源参考表
18	主数据参考信息	REF_INST_TYPE	机构类型参考表
19	主数据参考信息	REF_LEVEL	实体层级参考表

20	主数据参考信息	REF_EXCHANGE	交易所参考表

5. 投资研究主数据逻辑模型图

根据对投资研究主数据逻辑模型的数据分类，形成的逻辑模型图示例如图2所示，其相关数据表的详细设计内容见第7章。

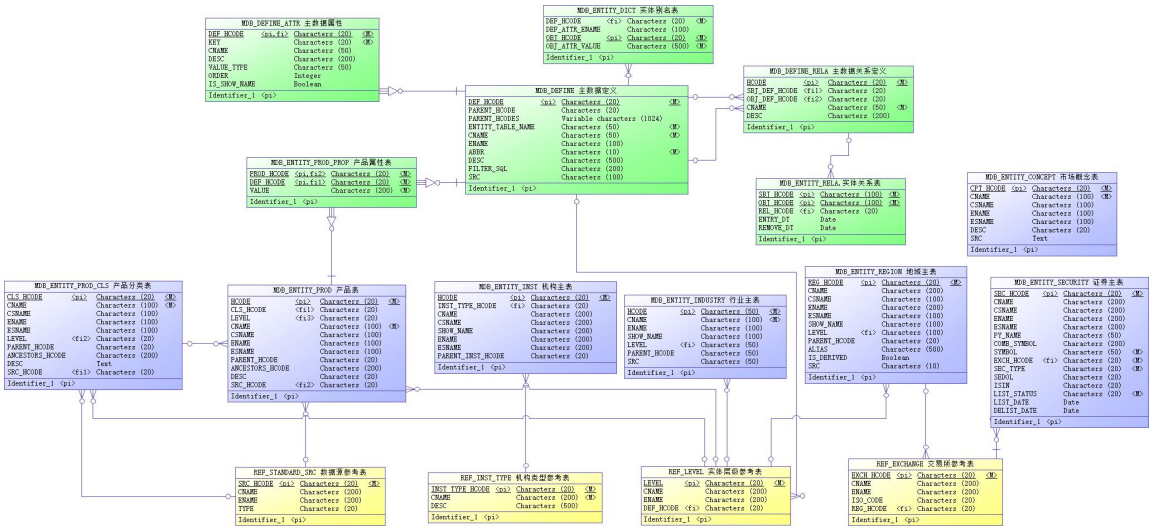


图 2 投资研究主数据逻辑模型图

6. 主数据模型数据表详细设计

通过梳理投资研究主数据逻辑模型的每个数据分类，分析整合形成出符合该分类定义范围的数据表信息，再对每个数据表进行细化设计，从而形成一套完整且适用于投资研究框架的实用性比较强的数据表结构。数据表之间通过相互关联，最终构成投资研究主数据逻辑模型的主体部分。

在投资研究主数据逻辑模型中，重要的数据主要为主数据定义与描述信息和主数据实体信息中的数据，该部分的数据表详细内容见6.1-6.16节。

其中，6.1-6.2章节所涉及到的表可划分为数据模型的定义表，6.3-6.9章节涉及到的表可被统称为业务表，6.10-6.14章节所涉及到的表为产品主表系列，产品数据在实际生活中品类繁多、内容更为丰富，每一品类具备其单独属性。但在下游日常的业务场景中，或者实际开发流程中，往往会统一调用。故针对产品主数据，我们未将其按不同的类别抽象成一个个独立的业务表，而是采用同一套数据模型进行维护，方便下游调用。

6.1 MDB_DEFINE 定义表

该表用于管理主数据，如新增业务、项目为主数据、更新已有内容。故在该表中需要维护主数据基本信息，如名称、业务描述、业务表名等信息，有统领主数据体系的作用，便于后续业务的查询。某一项主数据既可以作为业务本身，也可以作为描述其他业务的一个维度（或属性）。如地域，可以单独研究某一省市的地理、环境特征，也可以用来描述公司注册地、办公地。地域与公司皆被定义为主数据，内容见表2。

表 2 MDB_DEFINE 定义表

序号	英文字段	中文字段	字段类型	业务主键
1	HCODE	内部编码，英文，具备可读性。	varchar(20)	√
2	PARENT_HCODE	上级编码	varchar(20)	
3	PARENT_HCODES	祖先节点代码	json	
4	ENTITY_TABLE_NAME	所在业务表	varchar(50)	
5	NAME	主数据中文	varchar(50)	
6	ENAME	主数据英文	varchar(100)	
7	ABBR	用于生成下游业务表 HCODE 的前缀。	varchar(10)	
8	DESC	具体描述	varchar(500)	
9	FILTER_SQL	过滤条件	varchar(200)	
10	SRC	主要数据来源	varchar(100)	
注1：主数据可以有父子层级关系，如国家与省、直辖市。可以使用PARENT_HCODE、FILTER_SQL维护，PARENT_HCODES便于程序使用。 注2：FILTER_SQL 维护方式 WHERE和具体筛选条件。				

6.2 MDB_DEFINE_ATTR 主数据属性表

该表用于记录某一主数据具备的属性，如证券主数据需要有全称、简称、上市场所、上市日期等信息；地域需要维护中文名称、英文名称等内容，见表3。

表 3 MDB_DEFINE_ATTR 主数据属性表

序号	英文字段	中文字段	字段类型	业务主键
1	DEF_HCODE	内部编码	varchar(20)	√
2	KEY	字段英文名，主数据的属性	varchar(20)	√
3	NAME	字段中文名，主数据的属性	varchar(50)	
4	DESC	描述信息	varchar(200)	
5	VALUE_TYPE	数据类型	varchar(50)	
7	ORDER	排序	int(11)	
8	IS_SHOW_NAME	是否展示名称	tinyint(1)	

6.3 MDB_DEFINE_RELA 关系定义表

该表用于定义主数据之间的关系，如行业与证券的关系，公司与品牌的关系，内容见表4。

表 4 MDB_DEFINE_RELA 关系定义表

序号	英文字段	中文字段	字段类型	业务主键
1	HCODE	关系内部编码	varchar(20)	√
2	NAME	中文名	varchar(50)	
3	DESC	描述信息	varchar(200)	
4	SBJ_DEF_HCODE	主体，对应 MDB_DEFINE 的主数据内部编码	varchar(20)	
5	OBJ_DEF_HCODE	客体，对应 MDB_DEFINE 的主数据内部编码	varchar(20)	

6.4 MDB_ENTITY_INST 机构表

该表用于记录机构信息，包含中英文名称、中英文简称、办公地址、公司简介、经营范围、企业类型等业务信息。

该表的存在使得在数据库中的机构或组织有唯一的编码、有统一标准的信息。当数据库中其他业务表中出现了相关机构，通过程序或人工的方式将其匹配为机构表中的HCODE，实现机构数据标准化，内容见表5。

表 5 MDB_ENTITY_INST 机构表

序号	英文字段	中文字段	字段类型	业务主键
1	HCODE	内部代码	varchar(100)	√
2	CNAME	中文名称	varchar(200)	
3	CSNAME	中文简称	varchar(200)	
4	SHOW_NAME	中文拼音简称	varchar(200)	
5	ENAME	英文名称	varchar(200)	
6	ESNAME	英文简称	varchar(200)	
7	INST_TYPE_CD	机构类型 CODE	varchar(10)	
8	INST_TYPE	机构类型	varchar(50)	
9	PARENT_INST_HCODE	所属公司	varchar(100)	
10	REG_CAPITAL	注册资本	varchar(50)	
11	CURRENCY_UNIT	货币单位	varchar(20)	
12	COMPANY_STATUS	机构存续状态	varchar(20)	

13	ESTABLISHMENT_DATE	成立日期	varchar(20)	
14	CLOSE_DATE	存续截止日	varchar(20)	
15	REG_ADDR	注册地址	varchar(200)	
16	REG_COUNTRY	注册地所在国家	varchar(20)	
17	REG_CITY	注册所在省市	varchar(20)	
18	REG_AREA	注册所在区县	varchar(20)	
19	REG_ZIP	注册地址邮编	varchar(20)	
20	OFFICE_ADDR	办公地址	varchar(200)	
21	EMAIL	电子邮箱	varchar(50)	
22	WEBSITE	网址	varchar(50)	
23	TEL	联系电话	varchar(50)	
24	FAX	传真	varchar(50)	
25	BRIEF_INTRO_TEXT	公司简介	text	
26	BIZ_SCOPE	经营范围	text	
27	BUSINESS_MAJOR	主营业务	text	
28	WORKFORCE	员工人数	varchar(20)	
29	ORGANIZATION_CODE	组织机构代码	varchar(20)	
30	COMPANY_NATURE	企业性质	varchar(20)	
31	IS_LIST	是否上市	bit	
32	IS_BRANCH	是否分支机构	bit	
33	ORG_TYPE	组织形式	varchar(20)	
34	AUTH_CAPSK	法定股本	varchar(20)	
35	LEGAL_PERSON_REPR	法人代表	varchar(20)	
36	GENERAL_MANAGER	总经理	varchar(20)	
37	CHAIRMAN	董事长	varchar(20)	
38	BIZ_LICENSE_NO	营业执照号码	varchar(20)	
39	TAX_REGISTER_NO	税务登记证	varchar(20)	
40	ORG_REGISTER_NO	组织机构代码证	varchar(20)	

41	SOCIAL_CREDIT_NO	社会信用代码	varchar(20)	

6.5 MDB_ENTITY_SECURITY 证券表

该表用于记录各类证券信息，覆盖股票、债券、基金、指数等类型，包含证券全称、简称、上市地点、上市状态等业务信息，内容见表6。

表 6 MDB_ENTITY_SECURITY 证券表

序号	英文字段	中文字段	字段类型	业务主键
1	SEC_HCODE	证券内部代码	varchar(20)	√
2	CNAME	证券中文名称	varchar(200)	
3	COMB_SYMBOL	展示代码	varchar(200)	
4	EXCH_HCODE	交易所代码	varchar(200)	
5	CSNAME	证券中文简称	varchar(100)	
6	ENAME	证券英文名称	varchar(200)	
7	ESNAME	证券英文简称	varchar(100)	
8	PY_NAME	拼音缩写	varchar(50)	
9	SYMBOL	交易代码	varchar(50)	
10	TRADE_CURR_HCODE	交易货币代码	varchar(20)	
11	PARVALUE_CURR_HCODE	最小单位货币代码	varchar(20)	
12	SEC_TYPE	证券类型	varchar(20)	
13	SEDOL	SEDOL 编码	varchar(50)	
14	ISIN	ISIN 编码	varchar(50)	
15	LIST_STATUS	上市状态	varchar(20)	
16	LIST_DATE	上市日期	datetime	
17	DELIST_DATE	退市日期	datetime	
18	CSNAME_TRIM	交易简称	varchar(100)	

6.6 MDB_DEFINE 定义表

该表用于记录各类行业标准分类，覆盖申万等行业分类，内容见表7。

表 7 MDB_DEFINE 定义表

序号	英文字段	中文字段	字段类型	业务主键
1	HCODE	行业内部代码	varchar(50)	√
2	CODE	行业代码	varchar(50)	
3	CNAME	行业名称中文	varchar(100)	
4	ENAME	行业名称英文	varchar(100)	
5	SHOW_NAME	展示名称	varchar(100)	
6	LEVEL	行业分类层级代码	varchar(50)	
7	PARENT_HCODE	该条目父级代码	varchar(50)	
8	CODE_ALIAS	行业代码别名	varchar(50)	
9	SRC	行业分类标准出处	varchar(10)	

6.7 MDB_ENTITY_REGION 地域表

该表用于记录地区信息，覆盖大洲、国家、省、洲或直辖市、地级市等最细到街道的信息，内容见表8。

表 8 MDB_ENTITY_REGION 地域表

序号	英文字段	中文字段	字段类型	业务主键
1	HCODE	地域内部代码	varchar(50)	√
2	CODE	地域代码	varchar(50)	
3	CNAME	地域中文全称	varchar(200)	
4	CSNAME	地域中文简称	varchar(100)	
5	SHOW_NAME	展示名称	varchar(100)	
6	ENAME	地域英文全称	varchar(200)	
7	ESNAME	地域英文简称	varchar(100)	
8	PARENT_HCODE	上级编码	varchar(50)	
9	LEVEL	地域行政区划层级代码	varchar(100)	
10	ALIAS	地域别名	varchar(500)	
11	IS_DERIVED	是否为衍生地域	int(11)	
12	SRC	地域行政区划层级来源	varchar(10)	

6.8 MDB_ENTITY_PERSON 人物表

该表用于记录人物信息，包含人物名称、性别、出生年月、籍贯、教育背景等信息，内容见表9。

表 9 MDB_ENTITY_PERSON 人物表

序号	英文字段	中文字段	字段类型	业务主键
1	HCODE	人物内部代码	varchar(50)	√
2	CNAME	中文姓名	varchar(100)	
3	ENAME	英文姓名	varchar(45)	
4	SHOW_NAME	展示名称	varchar(45)	
5	BIRTHDAY	出生年月	varchar(20)	
6	SEX	性别	varchar(45)	
7	NATIONALITY	国籍	varchar(20)	
8	NATIVE_PLACE	籍贯	varchar(20)	
9	ETHNIC	民族	varchar(20)	
10	POLITICS	政治面貌	varchar(20)	
11	EDU	最高学历	varchar(100)	
12	MAJOR	最高学历所学专业	varchar(100)	
13	GRAD	最高学历毕业院校	varchar(100)	
14	COMP_HCODE	当前任职公司内部代码	varchar(500)	
15	TITLE	当前任职公司职位	varchar(200)	
16	NOTES	人物备注信息	varchar(300)	

6.9 MDB_ENTITY_CONCEPT 市场概念表

该表用于记录市场热点概念信息，如专精特新、一带一路、5G等信息，内容见表10。

表 10 MDB_ENTITY_CONCEPT 市场概念表

序号	英文字段	中文字段	字段类型	业务主键
1	HCODE	概念内部代码	varchar(50)	√
2	CODE	来源的概念代码	varchar(50)	
3	SOURCE_CODE	来源代码	varchar(50)	

4	CNAME	概念中文全称	varchar(512)	
5	CSNAME	概念中文简称	varchar(512)	
6	ENAME	概念英文全称	varchar(200)	
7	ESNAME	概念英文简称	varchar(100)	
8	IS_COM_CONCEPT	是否为综合概念	int(11)	
9	DES	概念描述	longtext	

6.10 MDB_ENTITY_PROD 产品表

产品表用于记录产品本身的信息，包含品类、名称，划分为品牌、系列、SKU三层进行维护，覆盖汽车、消费、医药等信息，内容见表11。

表 11 MDB_ENTITY_PROD 产品表

序号	英文字段	中文字段	字段类型	业务主键
1	HCODE	产品代码	varchar(50)	√
2	CLS_HCODE	品类代码，对应产品分类表	varchar(100)	
3	LEVEL_NAME	产品层级名称,P1-品牌，P2-系列，P3-产品	varchar(100)	
4	CNAME	产品中文全称	varchar(100)	
5	CSNAME	产品中文简称	varchar(100)	
6	SHOW_NAME	展示名称	varchar(100)	
7	ENAME	产品英文全称	varchar(100)	
8	ESNAME	产品英文简称	varchar(100)	
9	DESC	产品描述	varchar(100)	
10	PARENT_HCODE	父节点代码	text	
11	ANCESTORS_HCODE	祖先节点代码	varchar(100)	
12	SRC	来源	varchar(100)	

6.11 MDB_ENTITY_PROD_CLS 产品分类表

该表用于记录产品品类信息，可被视为产品分类主数据。与产品主表关联，获取同一个品类下的相关品牌、产品，内容见表12。

表 12 MDB_ENTITY_PROD_CLS 产品分类表

序号	英文字段	中文字段	字段类型	业务主键
1	CLS_HCODE	品类代码	varchar(100)	√
2	CODE	来源系统的代码	varchar(100)	
3	CNAME	品类中文名称	varchar(100)	
4	SHOW_NAME	展示名称	varchar(100)	
5	ENAME	品类英文名称	varchar(100)	
6	DESC	品类描述	text	
7	DESC_EN	品类英文描述	text	
8	PARENT_HCODE	品类父节点代码	varchar(100)	
9	ANCESTORS_HCODE	祖先节点代码	varchar(500)	
10	LEVEL_NAME	品类层级名称	varchar(100)	
11	SRC	来源	varchar(100)	

6.12 MDB_ENTITY_PROD_CLS_ATTR 品类属性表

该表用于记录某一产品品类下的关键属性，是MDB_DEFINE和MDB_ENTITY_PROD_CLS两表关联。通过产品表和产品属性值表聚合成某一品类的属性集合，内容见表13。

表 13 MDB_ENTITY_PROD_CLS_ATTR 品类属性表

序号	英文字段	中文字段	字段类型	业务主键
1	CLS_HCODE	品类代码	varchar(100)	√
2	DIM_HCODE	品类的属性,对应概念定义表的概念代码	varchar(100)	√

6.13 MDB_ENTITY_PROD_PROP 产品属性值表

该表用于记录产品的属性，如可以获取到汽车的车型、动力类型、级别等内容。该表以窄表方式存储，内容见表14。

表 14 MDB_ENTITY_PROD_PROP 产品属性值表

序号	英文字段	中文字段	字段类型	业务主键
1	P_HCODE	产品代码，对应产品主表中的产品代码	varchar(100)	√
2	DIM_HCODE	产品的属性，对应定义表的内部编码	varchar(100)	√
3	VALUE	对应的属性值	varchar(100)	

6.14 MDB_ENTITY_DICT 主数据实体别名表

该表用于记录各类主数据的别名，获取标准名称和别名的关系。通过品类与产品，产品与属性表，可以归纳出某一品类下具备的属性，内容见表15。

表 15 MDB_ENTITY_DICT 主数据实体别名表

序号	英文字段	中文字段	字段类型	业务主键
1	DEF_HCODE	主数据类型，对应 MDB_DEFINE 定义代码	varchar(20)	
2	DEF_ATTR_ENAME	属性名	varchar(20)	√
3	OBJ_HCODE	对应的实体	varchar(200)	
4	OBJ_ATTR_VALUE	别名	varchar(250)	

6.15 MDB_ENTITY_COMMON 通用实体表

该表记录关注度较低、当前业务中短期暂时不对其进行专门研究、暂不关注其特有属性的主数据，可用一张通用表记录，内容见表16。

表 16 MDB_ENTITY_COMMON 通用实体表

序号	英文字段	中文字段	字段类型	业务主键
1	DEF_HCODE	主数据类型，对应 MDB_DEFINE 定义代码	varchar(20)	
2	HCODE	内部编码	varchar(20)	√
3	SHOW_NAME	展示名称	varchar(20)	
4	PARENT_HCODE	父节点的 HCODE	varchar(20)	

6.16 MDB_ENTITY_COMMON 通用实体表

该表用于记录不同主数据之间的关系，内容见表17。

表 17 MDB_ENTITY_COMMON 通用实体表

序号	英文字段	中文字段	字段类型	业务主键
1	REL_HCODE	关系的类型	varchar(20)	√
2	SBJ_HCODE	主体实体的内部编码	varchar(100)	
3	OBJ_HCODE	客体实体的内部编码	varchar(100)	
4	ENTRY_DT	生效时间	date	
5	REMOVE_DT	失效时间	date	

7. 投资研究主数据逻辑模型应用场景

7.1 场景一 数据表意消歧

构建完整的投资研究主数据模型，有助于消除以下场景中的表意歧义，内容见表18。

表 18 产品编码表

产品编码	中文名称	品类
HPRD0000000000000003	苹果	农产品
HPRD0000000000000013	苹果	电子产品品牌

如在下游指标“苹果:销售额”中，若无其他业务场景辅助判断，则对“苹果”字段会存在歧义。利用主数据模型可以产品编码“HPRD0000000000000013” 快速定位到目标，提高对接的效率。

7.2 场景二 指标表达标准化

构建完整的投资研究主数据模型，有助于促进指标表达的标准化。如表19记录了由于相同内容不同数据商的不同表达而造成的数据干扰。

表 19 不同供应商指标对比表

供应商 指标	供应商 W	供应商 T	供应商 R
指标 1	产量:家用电冰箱:安徽:当月值	安徽:规模以上工业产品产量:家用电冰箱:当月值	家用电冰箱产量:当月值:安徽省
指标 2	出厂价:聚合 MDI:上海巴斯夫	出厂价: MDI(聚合):上海巴斯夫	出厂价:聚氨酯:MDI(聚合):上海巴斯夫
指标 3	销量:汽车:A 级:一汽大众:捷达:当月值	汽车销量:一汽大众: 大众捷达:当月值	乘用车销量:大众捷达:累计值

以上所列情况中包含以下几个共性问题：

a) 实体命名不规范（安徽/安徽省、聚合 MDI/MDI(聚合)、一汽大众:捷达/一汽大众: 大众捷达/大众捷达）；
b) 指标命名不规范（销量:轿车/乘用车销量/汽车销量）；
c) 维度顺序不一致。

解决如上问题可以做以下标准化内容：

a) 地域主数据中，HREG000000001015 对应安徽省，简称为安徽。给指标 1 中的安徽匹配标准编码；
b) 产品主数据中，HPRD00000072484 对应聚合 MDI,别名为粗 MDI、聚氨酯黑料、MDI(聚合)。给指标 2 中的聚合 MDI 匹配标准代码；
c) 产品主数据中，HPRD0000000000001343 对应捷达，别名为一汽大众捷达。给指标 3 中的捷达匹配标准代码；
d) 产品品类表中，HCLS0000000000001592 对应家用冰箱。给指标 1 中的家用电冰箱匹配标准代码；

e） 机构主数据中，HINST0000000000001972 对应上海巴斯夫聚氨酯有限公司，简称为上海巴斯夫，给指标 2 匹配上海巴斯夫的标准代码；HINST0000000000001973 对应一汽大众汽车有限公司，简称为一汽大众，给指标 3 一汽大众匹配标准代码。

f） 另将统计口径、指标标准化处理后编码。

通过指标表达的标准化，最终成果如表 20 所示：

表 20 指标标准化成果表

指标	原指标名称	指标维度	维度编码	维度中文
指标 1	1.产量:家用电冰箱:安徽:当月值 2.安徽:规模以上工业产品产量:家用电冰箱:当月值 3.家用电冰箱产量:当月值:安徽省	指标	HABS00001000051	产量
		统计口径	HIDTP00000000001	当月值
		品类	HCLS000000000001592	家用冰箱
		地域	HREG00000001015	安徽
指标 2	1.出厂价:聚合 MDI:上海巴斯夫 2.出厂价: MDI(聚合):上海巴斯夫 3.出厂价:聚氨酯:MDI(聚合):上海巴斯夫	指标	HABS00001002849	出厂价
		产品	HPRD00000072484	聚合 MDI
		公司	HINST0000000000001972	上海巴斯夫
指标 3	1.销量:汽车:A 级:一汽大众:捷达:当月值 2.汽车销量:一汽大众: 大众捷达:当月值 3.乘用车销量:大众捷达:累计值	指标	HABS00000000327	汽车销量
		统计口径	HIDTP00000000001、HIDTP00000000002	当月值、累计值
		公司	HINST0000000000001973	一汽大众
		品牌	HPRD000000000001343	捷达

经过标准化处理后，各指标由于被拆解为各类标准维度，能方便下游快速地确认指标一致性，同时也实现了标准化展示：

g） 指标 1：产量:家用冰箱:安徽:当月值

h） 指标 2：出厂价:聚合 MDI:上海巴斯夫

i） 指标 3：汽车销量:一汽大众:捷达:当月值、汽车销量:一汽大众:捷达:累计值

7.3 场景三 数据智能应用

为保障投研信息及时有效地传递，数据的有效筛选和标识是重要影响因素之一。主数据实现了各类实体的标签以及实体之间的关系，如图3所示。

图 3 实体标签及实体之间关系图

采用投资研究主数据模型各数据表中的关联关系，可通过搜索“海信 冰箱”搜索到“容声”冰箱的相关指标，搜索结果如图4所示。

图 4 指标关联搜索结果图

8. 指标标准模型设计方案

8.1 背景

在当前的投资研究业内，常用的指标表示方式，有个约定俗成的方式，即使用冒号将原子指标的名称和不同维度的取值拼接形成展示名称。例如：

北京:GDP:第二产业:工业:累计值

在上述标准模型下，“北京:GDP:第二产业:工业:累计值”是一个派生指标，其对应的原子指标为“GDP 累计值”。对于度量“GDP 累计值”来说，有两个描述该度量的维度：地域、产业类型。

遗憾的是，当前的投资数据流通领域，数据的表达方式通常为一个时间序列的派生指标，但是派生指标由哪个原子指标而生，每个冒号间的词语是什么含义（例如“北京”是个地区还是个公司？），

并不在数据流通的清单之内。这样的数据结构，只能用于场景单一的时序数据展示。无论是分析师进行多指标、多维度的复杂分析，还是计算机进行联机数据分析或人工智能计算，都是远远不够的。

在人工智能和大数据开始渗透到金融行业的每一个角落，投资研究领域也不例外。“人工智能”的基石就是让机器具备足够的知识。相比于标准化程度较高的传统金融市场资讯数据来说，宏观经济、垂直行业和领域数据非标准化程度高，机器学习的难度相应较大。因此一套通用的能够表示投研数据指标多维度信息的标准化数据模型，无论对于复杂数据的传输共享还是机器学习与智能，都是非常必要的。

8.2 标准化模型设计

原子指标和派生指标的关系，可以用经典的关系代数模型表示：

$$Y_{Measure} = f(t, x1, x2, x3, \dots, xn)$$

Y：派生指标数值；

X：入参向量，即“维度”。常见的派生指标为时间序列数据，因此用 t 表示时间入参。其余入参用$x1, x2, x3, \dots, xn$表示。非时间维度还可区分为标识性维度和非标识性维度；

f：度量，即原子指标，也是获取派生指标数值的函数，指定入参向量值（即维度值）后返回确定的数据集。

在上例中，x1 = 北京，x2 = 工业，如果再取 t = 2019-12-31，则可获得 2019 年末北京在工业的 GDP 全年累计值。如果不传入 t 参数，则可获取北京自有统计数据以来的工业 GDP 的全部时间序列数据

值得注意的是，在原始指标名称中，“第二产业”并不是个标识性维度，它只是用来描述“工业”，方便使用指标的人理解而加上的。在$Y = f(t, x)$的模型中，“第二产业”并不是决定 Y 值的变量。

本文基于经典的关系代数理论，制定了一套通用的指标标准化模型，用以在提供派生指标时间序列数值的基础上，完整、标准地定义指标的维度信息，供分析师和机器更灵活地消费指标数据。下文详述模型中每张表的设计及相关规范。

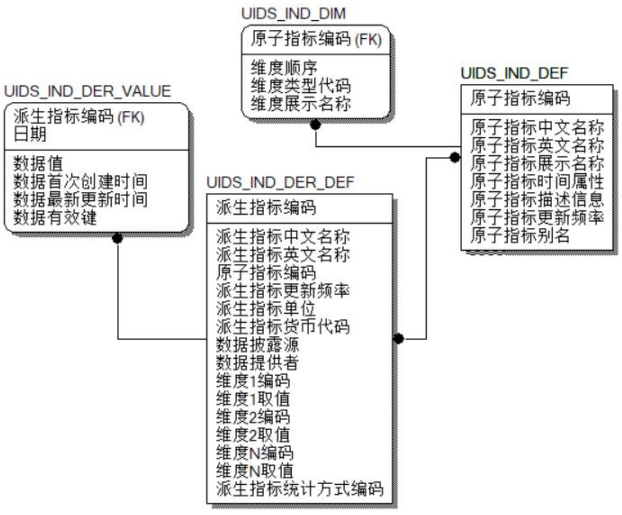


图 5 投资研究标准化数据逻辑模型

8.2.1 UIDS_IND_DEF（原子指标定义表）

该表用于定义原子指标。原子指标是一类拥有共同维度类型且业务含义、统计口径一致的指标集合。原子指标的定义方式取决于生数据的存储形态。如果生数据通过关系型数据库存储，通常来说，一个原子指标来自于一张表或视图的一个字段，且同一个数据集的原子指标拥有相同的维度定义；如果生数据已经是派生指标形态，则需要根据业务知识抽象的方式“还原”原子指标的名称，例如：“江苏:工业增加值:累计值”和“湖南:工业增加值:累计值”的原子指标为“工业增加值”。

该表需要定义原子指标的标准名称和展示名称。二者的区别在于标准名称面向数据管理人员，展示名称面向业务分析人员。例如“市盈率”是个标准原子指标名称，但是对于业务人员来说，“PE”是个更习惯的展示名称。同时，如果可能需尽量穷尽原子指标的多种不同名称，以便数据作业任务能够识别同义不同名的原子指标。

其他诸如原子指标的频率、来源、时间属性等，也需数据管理人员进行明确定义。

表 21 原子指标定义表

表名称	数据项名称	数据项类型	详细定义	数据取值范围	主键
UIDS_IND_DEF 原子指标定义表	原子指标编码	String	原子指标的编码		Y
	原子指标中文名称	String	原子指标标准中文名称		
	原子指标英文名称	String	原子指标标准英文名称		
	原子指标展示名称	String	原子指标面向用户展示的名称		
	原子指标时间属性	String	指标的统计方式是时点或区间	时点：spot，区间：duration	
	原子指标描述信息	String	原子指标的详细描述，如计算方法，适用范围等		
	原子指标更新频率	String	原子指标的更新频率	Y / HY / Q / M / W / D / RT	
	原子指标别名	String	原子指标的别名列表，用“，”分隔不同别名		

8.2.2 UIDS_IND_DIM（原子指标维度表）

该表用于定义原子指标的维度，也是机器能够理解指标业务含义的重要信息来源。通常来说，时间维度作为所有投资指标中必不可少的维度，不必在该表的每个指标定义中单独列示。

实际业务中，不同经济领域和行业中的原子指标维度各异，有的只有一个维度，有的则有五个甚至以上的维度，因此采取纵表的形式存储原子指标的维度信息。其中，“维度顺序”字段使用自

增整数来定义每个维度之间的次序关系。该次序需根据常识和业务规则制定。在未来根据不同的维度值拼装时，维度顺序被用来按序拼接指标名称。例如原子指标“GDP 累计值”中，地域维度顺序为 1，产业类型维度顺序为 2，则其派生指标名称应定义为“北京:GDP:第二产业:工业:累计值”。

表 22 原子指标维度表

表名称	数据项名称	数据项类型	详细定义	数据取值范围	主键
UIDS_IND_DIM 原子指标维度表	原子指标编码	String	原子指标的编码		Y
	维度顺序	Int	维度按照业务逻辑的排列	1,2,3,…,N	
	维度类型编码	String	根据维度的值域归纳的维度类型编码	region, security, product, person,…	
	维度展示名称	String	根据维度的值域归纳的维度类型的展示名称	地域、证券、产品、人…	

8.2.3 UIDS_IND_DER_DEF（派生指标定义表）

该表用于定义派生指标的信息。派生指标的定义与原子指标定义相比，需要额外定义原子指标在每个维度中的枚举值。例如对于“北京:GDP:第二产业:工业:累计值”，维度 1 取值为“北京”，维度 2 取值为“工业”。

与此同时，需要为派生指标定义其所属的原子指标编码。

表 23 派生指标定义表

表名称	数据项名称	数据项类型	详细定义	数据取值范围	主键
UIDS_IND_DER_DEF 派生指标定义表	派生指标编码	String	派生指标的编码		Y
	派生指标中文名称	String	派生指标中文名称		
	派生指标英文名称	String	派生指标英文名称		
	原子指标编码	String	原子指标的编码		
	派生指标单位	String	派生指标的单位/量纲		
	派生指标货币编码	String	派生指标的货币编码	CNY, USD, EUR, …	
	派生指标更新频率	String	派生指标的更新频率	Y / HY / Q / M / W / D / RT	
	数据披露源	String	派生指标的原始披露来源		
	数据提供者	String	派生指标的数据提供者		

	维度 1 取值	String	派生指标所属原子指标的维度 1 取值		
	维度 2 取值	String	派生指标所属原子指标的维度 2 取值		
	维度 N 取值	String	派生指标所属原子指标的维度 N 取值		
	派生指标统计方式	String	派生指标的统计方式编码	CURR, AGG, YOY, …	

8.2.4 UIDS_IND_DER_VALUE（派生指标数值表）

在投资研究的业务场景中，指标通常以时间序列的形态展示。该表以派生指标代码和日期为联合主键，专用于列示派生指标的取值时间序列。

从工程层面，该表需制定三个系统键值：数值有效键、数据首次创建时间、数据最新更新时间。其中该表采取逻辑删除机制。所有记录的数值有效键默认值为 True，即若历史数据发生删除，也不执行物理删除动作，而将该键设置为 False。两个系统时间戳除去为业务用户提供数据更新状态的参照，也为上下游 ETL 任务提供灵活性。

表 24 派生指标数值表

表名称	数据项名称	数据项类型	详细定义	数据取值范围	主键
UIDS_IND_DER_VALUE 派生指标数值表	派生指标编码	String	派生指标编码		Y
	日期	Date	数据的日期点		Y
	数值	Number	派生指标在该日期的取值		
	数值有效键	Bool	该记录当前是否有效	T: 有效; F: 无效	
	数据首次创建时间	Datetime	该记录首次进入系统的时间戳		
	数据最新更新时间	Datetime	该记录最新在系统中更新的时间戳		

8.3 派生指标命名规范

根据前文所述，派生指标的名称由以下元素拼接而成：

- 维度 1，维度 2,…，维度 N 的码值
- 原子指标名称
- 数据频率
- 统计方式与口径

鉴于当前投研数据应用环境中，业务人员已经习惯于用“：”拼接不同实体取值的表达方式，本文对于派生指标名称的拼接顺序约定如下：

{维度 1 取值}:{维度 2 取值}:…:{维度 N 取值}:{原子指标名}:{统计方式与口径}:{频率}

以上文提到的指标“北京:GDP:工业:累计值”为例，其标准化的指标名称表达为：

北京:工业:GDP:累计值:季

其中，“北京”为维度 1“地域”的取值，“工业”为维度 2“产业类型”的取值，“GDP”为原子指标名称，“累计值”为指标的统计方式，“季”为数据的更新频率。

9 维度表达标准化设计方案

对于指标类数据，物理模型的统一标准是必要但不充分的。为让数据能够自由流通且让机器尽量自动化地学习数据背后的业务含义，维度类型和维度枚举值的标准化也是至关重要的。投资研究行业的业务模型相对于传统金融业务（如基金管理、证券承销、市场营销等），要覆盖全部的宏观和行业领域，要复杂得多。不同行业的指标，都拥有该行业特性的维度属性。例如汽车行业：车型、车系、排量等；新能源电池领域则有：直径、电芯种类、包装类型等。因此，从经济生产领域归纳、抽象一些通用的维度和实体并形成通用准则，是实现数据标准化的重要一环。

9.1 维度类型编码（主数据）规范

“3.2 实体”一节对投资研究领域的一些通用实体做了归纳和定义。对于这些实体类型，规范定义了通用维度类型的中文和英文及英文词根的标准表达方式，如表 25 所示：

表 25 维度类型标准名称定义表

序号	中文全称	英文全称	编码
1	证券	Security	sec
1a	股票	Stock	stk
1b	债券	Bond	bond
2	基金	Fund	fund
3	期货	Futures	futr
4	指数	Index	idx
5	公司	Organization	org
6	地域	Region	reg
7	人物	Person	psn
8	产品	Product	prod
9	渠道	Channel	chnl
10	行业	Industry	indu

其中，证券下属的不同产品类型的命名规范，沿用《证券期货业数据模型 第 4 部分：基金公司逻辑模型》[2]。

9.2 通用维度码值规范

每一种通用维度在不同行业的表示方式均有不同，为避免同义不同名的表述方式对数据的流通造成的障碍，对上述通用维度的编码和枚举值进行规范。

9.2.1 证券及相关金融工具 security

全球证券使用 ISO 6166《证券及相关金融工具－国际证券识别编码体系》中规定的全球证券唯一识别码（ISIN）。示例：

表 26 证券实体标准化码值表

ISIN	证券代码	中文全称	中文简称	证券类型
CNE100001QW3	06818.HK	中国光大银行股份有限公司	中国光大银行	STK
KYG8972T1067	01083.HK	港华燃气有限公司	港华燃气	STK
CNE100000288	08247.HK	中生北控生物科技股份有限公司	中生北控生物科技	STK
KYG215AR1066	01492.HK	中国中地乳业控股有限公司	中地乳业	STK
KYG9716W1087	03322.HK	永嘉集团控股有限公司	永嘉集团	STK
KYG5138B1023	02362.HK	金川集团国际资源有限公司	金川国际	STK
BMG7946B1000	00251.HK	爪哇控股有限公司	爪哇控股	STK

表 27 证券实体标准化码值表

ISIN	证券代码	中文全称	中文简称	证券类型
CNE100000VH6	150021.SZ	富国汇利回报分级债券型证券投资基金之 B 份额	富国汇利 B	FUND
CNE100000VH6	150021J.SZ	富国汇利分级债券型证券投资基金 B 级	富国汇利 B	FUND
CNE1000017M9	150022.SZ	申万菱信深证成指分级证券投资基金申万收益份额	申万深成指 A	FUND
CNE1000017L1	150023.SZ	申万菱信深证成指分级证券投资基金申万进取份额	申万深成指 B	FUND
CNE100000W29	150025.SZ	大成景丰分级债券型证券投资基金 A 类份额	大成景丰 A	FUND
CNE100000W37	150026.SZ	大成景丰分级债券型证券投资基金 B 类份额	大成景丰 B	FUND
CNE100000XW1	150027.SZ	天弘添利分级债券型证券投资基金 B 级	天弘添利 B	FUND

表 28 证券实体标准化码值表

ISIN	证券代码	中文全称	中文简称	证券类型
N/A	IH1805.CFFEX	上证 50 股指期货 1805 合约	上证 50 股指期货 1805	FUT
N/A	IH1804.CFFEX	上证 50 股指期货 1804 合约	上证 50 股指期货 1804	FUT
N/A	T1903.CFFEX	10 年期国债期货 1903 合约	10 年期国债期货 1903	FUT
N/A	IC1804.CFFEX	中证 500 股指期货 1804 合约	中证 500 股指期货 1804	FUT
N/A	IH1808.CFFEX	上证 50 股指期货 1808 合约	上证 50 股指期货 1808	FUT
N/A	IF1805.CFFEX	沪深 300 股指期货 1805 合约	沪深 300 股指期货 1805	FUT
N/A	IC1903.CFFEX	中证 500 股指期货 1903 合约	中证 500 股指期货 1903	FUT

9.2.2 公司 organization

采用企业在工商登记信息中记录的统一社会信用代码。示例：

表 29 公司实体标准化码值表

统一社会信用代码	公司全称	公司简称	注册省份	注册市县	法人代表
9132090273531238X1	江苏同隆建设集团有限公司	江苏同隆建设集团	江苏省	盐城市	陈同文
91320982695522007D	江苏大丰中洲建设工程有限公司	大丰中洲建设	江苏省	盐城市	王永进
915101126675584324	成都高原汽车工业有限公司	高原汽车	四川省	成都市	安聪慧
913209821406641084	江苏东远建筑有限公司	东远建筑	江苏省	盐城市	胡永东
913207071392576016	江苏苏港工程有限公司	苏港工程	江苏省	连云港市	陈振房
91320707567762778G	江苏伟仁建设工程有限公司	伟仁建设	江苏省	连云港市	仲伟仁
91320923678955863U	江苏恒健建设集团有限公司	江苏恒健建设集团	江苏省	盐城市	孙荣
9150010468145893XR	重庆明峰水处理设备有限公司	重庆明峰水处理设备	重庆	重庆市	李恒伟
913205096083033417	江苏姑苏净化科技有限公司	江苏姑苏净化	江苏省	苏州市	章洪伟

91320684750503512L	苏通建设集团有限公司	苏通建设集团	江苏省	南通市	蔡国新
913206121387436080	南通英雄建设集团有限公司	英雄建设	江苏省	南通市	沈锋

9.2.3 地域 region

国内地域采用中华人民共和国民政部《县以上（下）行政区划代码》。

国际地域因类型不同而标准不同。国家级地区使用 ISO 3166-1:2006（International Standard Norme Internationale）[3]中规定的全球国家和地区的三字母代码；国家以下地区遵循各国家的民政部门的统一命名和编码标准。示例：

表 30 地域实体标准化码值表

地域代码	中文名称	英文名称	地域等级
110000	北京市		省、洲或直辖市
110101	东城区		地级市
110102	西城区		地级市
710000	台湾省		省、洲或直辖市
SAM	南美洲	South America	大洲
OCE	大洋洲	Oceania	大洲
EUR	欧洲	Europe	大洲
USA	美利坚合众国	United States of America	国家
10001	纽约市	New York City	地级市

9.2.4 人物 person

因人物的统一编码身份证号不属于社会公开信息，人物的编码由各数据供给方独立采编。

采编原则：使用姓名、性别、任职公司为联合主键哈希生成。

9.2.5 产品 product

使用国家统计局《统计用产品分类目录》2010 年版[4]。示例：

表 31 产品实体标准化码值表

统计局产品编码	类别	产品名称
1	大类	农业产品
101	中类	谷物
10101	小类	稻谷
1010101	组	早籼稻
101010101	小组	种用早籼稻
101010199	小组	其他早籼稻
1010102	组	晚籼稻
101010201	小组	种用晚籼稻
101010299	小组	其他晚籼稻
1010103	组	中籼稻

9.2.6 行业 industry

使用中华人民共和国金融行业标准《上市公司分类与代码》（文档编号 JR/T 0020—2004）中定义的国民经济行业分类。示例：

表 32 行业实体标准化码值表

行业编码	类别名称	父类行业	行业层级
A	农、林、牧、渔业		大类
A01	农业	A	中类
A0101	谷物及其他作物种植业	A01	小类
A0120	蔬菜、园艺作物种植业	A01	小类
A0130	水果、坚果、饮料及香料作物种植业	A01	小类
A0150	中药材种植业	A01	小类
A03	林业	A	中类

9.2.7 渠道 channel

渠道通常用来分析上市公司发行产品的收入来源分布情况。本文归纳了 10 余种常见的不同行业的销售渠道并进行编码，后续可根据各行业需要持续扩充。

表 33 渠道实体标准化码值表

渠道编码	渠道中文名称	渠道英文名称
CH001	居家购物	Homeshopping
CH002	电子商务	E-Commerce
CH003	直销	Direct Selling
CH004	杂货店	Grocery Retailers
CH005	其他非杂货专营	Other Non-Grocery Specialists
CH006	户外市场	Outdoor Markets
CH007	百货商店	Department Stores
CH008	大卖场	Mass Merchandisers
CH009	仓储式商店	Warehouse Clubs
CH010	超市	Supermarkets
CH011	线下经销商	Online Dealer
CH012	线上经销商	Offline Dealer
CH013	银行	Banks
CH014	券商	Brokers

9.2.8 币种 currency

币种在金融数据中是一个常见的字段，也是一种特殊维度。在上述通用指标模型中，币种和时间维度一样，被独立成标准字段。币种的码值规范使用 ISO 4217（CURRENCY CODES）中规定的三字母货币代码。示例：

表 34 币种实体标准化码值表

ISO 货币代码	货币中文名称	货币英文名称
CNY	人民币元	China Renminbi
ARS	阿根廷比索	Argentine Peso
ATS	奥地利先令	Austrian Schilling
AUD	澳大利亚元	Australian Dollar
BDT	孟加拉塔卡	Bangladesh Taka
BEF	比利时法郎	Belgian Franc
BRL	巴西里亚尔	Brazil Real
CAD	加拿大元	Canadian Dollar
CHF	瑞士法郎	Swiss Franc
CNH	离岸人民币元	Offshore RMB
EUR	欧元	EURO

派生指标所有涉及的通用维度的取值和编码，均应按照以上规范在 UIDS_IND_DER_DEF（派生指标定义表）表中赋值。如遇到上述规范无法覆盖的维度码值，需在“维度 N 编码”字段添加“N/A”标识。

9.2.9 统计方式 calc_prop

时序指标通常有统计方式和口径的概念，常见的统计方式如同比、环比、累计同（环）比等。对于同一指标概念的不同统计方式下的数值，本文认为应统一为同一原子指标，用统计方式属性的不同编码来进行区分标识。常用的统计方式编码建议如下表：

表 35 统计方式属性标准化码值表

统计方式编码	统计方式中文名称	统计方式别名
CURR	当期值	当月值,当周值,当季值
AGG	累计值	月累计,周累计,季累计
YTD	年度至今累计值	YTD
YOY	同比	同比值,同比增速
MOM	环比	环比值,环比增速
AGG_YOY	累计同比	累计同比增速
AGG_MOM	累计环比	累计环比增速

10 通用指标模型 ETL 作业的设计方法

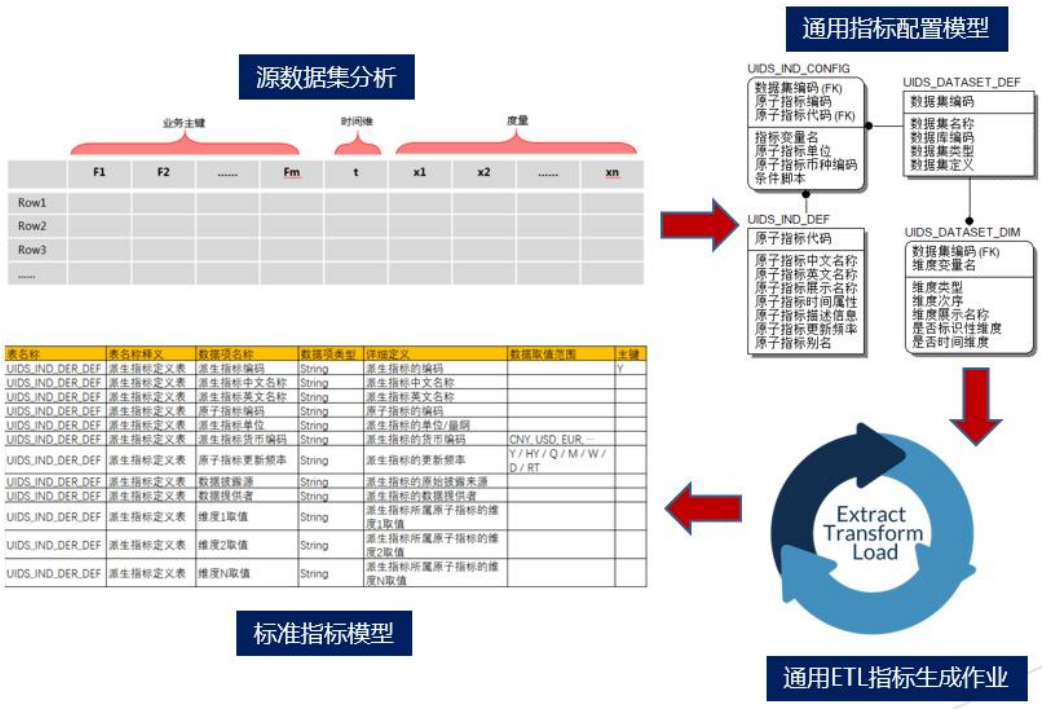
10.1 概述

根据投研数据应用实践总结，宏观经济领域的派生指标量通常在 40-80 万，各行业级垂直数据的指标量则在百万级。数百万级的派生指标按照上述标准逻辑模型，由业务人员或者数据管理人员定义，工作量是巨大的。因此，需要一套标准化的数据作业流程，使得数据管理人员在尽量少且标准化的配置工作后，由机器作业，生成标准化的派生指标。

10.2 方法简述

在标准指标模型 $Y_{Measure} = f(t, x1, x2, x3, \ldots, xn)$ 中，可知派生指标由原子指标根据各维度的参数派生而来，因此派生指标的维度定义可继承自原子指标。进一步分析，在关系型数据库表中，同一源数据集的不同度量（也即原子指标）共享同样的维度定义。因此，只要数据管理人员可以标准化地定义数据集、数据集的原子指标和数据集的维度信息，原子指标的维度、派生指标的维度和码值均可由数据作业自动生产。

源数据集定义：源数据集可以是关系型数据库表，也可以是视图，或标准化的 API、json、csv 文件，也即任何可以用二维关系表示的数据结构包含的数据集合。



表名称	表名称释义	数据项名称	数据项类型	详细定义	数据取值范围	主键
UIDS_IND_DER_DEF	派生指标定义表	派生指标编码	String	派生指标的编码		Y
UIDS_IND_DER_DEF	派生指标定义表	派生指标中文名称	String	派生指标中文名称		
UIDS_IND_DER_DEF	派生指标定义表	派生指标英文名称	String	派生指标英文名称		
UIDS_IND_DER_DEF	派生指标定义表	原子指标编码	String	原子指标的编码		
UIDS_IND_DER_DEF	派生指标定义表	派生指标单位	String	派生指标的单位/量纲		
UIDS_IND_DER_DEF	派生指标定义表	派生指标货币编码	String	派生指标的货币编码	CNY, USD, EUR, ...	
UIDS_IND_DER_DEF	派生指标定义表	原子指标更新频率	String	派生指标的更新频率	Y / HY / Q / M / W / D / RT	
UIDS_IND_DER_DEF	派生指标定义表	数据披露源	String	派生指标的原始披露来源		
UIDS_IND_DER_DEF	派生指标定义表	数据提供者	String	派生指标的数据提供者		
UIDS_IND_DER_DEF	派生指标定义表	维度1取值	String	派生指标所属原子指标的维度1取值		
UIDS_IND_DER_DEF	派生指标定义表	维度2取值	String	派生指标所属原子指标的维度2取值		
UIDS_IND_DER_DEF	派生指标定义表	维度N取值	String	派生指标所属原子指标的维度N取值		

图 6 通用指标模型 ETL 作业流程示意图

- 源数据集提炼。源数据集可以是一张关系型数据库表，也可以是一个视图，或 API/json/csv。数据管理人员需在源数据集根据数据字典和业务知识，归纳得出业务主键（标识性维度、非表示性维度），时间维度，度量（原子指标）；
- 通用指标模型配置。将抽象出的信息分别配置在源数据集定义、源数据集维度、原子指标定义、原子指标配置表中；
- ETL 作业读取通用配置表和源数据集数据，生成派生指标；

- 派生指标按照派生指标标准模型写入数据存储介质。

10.3 通用指标配置模型详述

通用指标模型生成作业的核心是第二步“通用指标模型配置”。将多源、异构的生数据通过数据管理人员的归纳，用统一的模型进行描述，是后续 ETL 程序批量处理和生成标准化指标的基石。下面详述每一张配置表的核心内容。

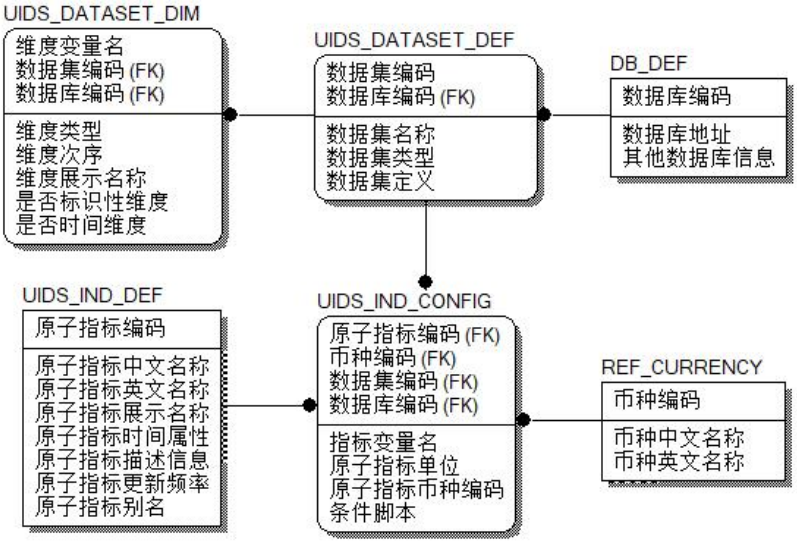


图 7 通用指标配置逻辑模型

10.3.1 UIDS_DATASET_DEF（源数据集定义表）

该表以源数据集为单位，定义一批拥有相同维度信息的原子指标的集合。该表的“数据库编码”字段尚需关联数据库配置表以获取源数据集的接入方式：若为数据库，则为数据库的连接串；若为 csv 或 json 文件，则是文件的存储路径。该表不是本模型的核心表故省略。

需要注意的是，无论源数据集的物理形态如何，在源数据集定义字段需制定源数据集的更新时间戳 updatetime，供 ETL 作业增量抽取数据。

表 36 源数据集定义表

表名称	数据项名称	数据项类型	详细定义	数据取值范围	主键
UIDS_DATASET_DEF 源数据集定义表	源数据集编码	String	源数据集的编码		Y
	源数据集名称	String	源数据集的标准名称		
	数据库编码	String	源数据集所在数据库的编码		
	源数据集类型	String	源数据集的物理形态	sqlscript / json / csv	
	源数据集定义	Text	获取源数据集数据的方式（如 SQL 脚本）	SQL 脚本 / python 脚本	

10.3.2 UIDS_IND_DEF（原子指标定义表）

同 3.2.1“UIDS_IND_DEF（原子指标定义表）”。

10.3.3 UIDS_IND_CONFIG（原子指标配置表）

原子指标于数据集为多对多的关系，也即一个原子指标可能出现在不同的数据集中。该表配置原子指标“在哪个数据集（源数据集编码）”，“通过哪个字段获取（原子指标变量名）”。条件脚本字段在必要时用以和源数据集定义表的“数据集定义”字段拼接，给予 ETL 程序更高的灵活性。例如，当源数据的形态为纵表，原子指标按 var 列纵向排列时，可将条件脚本置为 where var = ‘sample value’供 ETL 程序从源数据集读取。

表 37 原子指标配置表

表名称	数据项名称	数据项类型	详细定义	数据取值范围	主键
UIDS_IND_CONFIG 原子指标配置表	数据集编码	String	源数据集的编码		Y
	数据库编码	String	源数据集所在数据库的编码		
	原子指标编码	String	原子指标的编码		Y
	指标变量名	String	原子指标对应源数据集的字段名称，供 ETL 程序读取		Y
	原子指标单位	String	原子指标的单位/量纲		
	原子指标币种编码	String	原子指标的币种编码（如需）	CNY, USD, EUR, …	
	条件脚本	String	拼接原子指标的数据读取脚本（如需）		

10.3.4 UIDS_DATASET_DIM（源数据集维度定义表）

该表定义每个数据集的维度信息，也即“哪个维度（维度类型）在哪个数据集（源数据集编码）通过哪个字段（维度变量名）取值”。“维度次序”的定义在必要时取决于业务逻辑，例如“苹果:手机:销售额:当月值”这个指标，“品牌维度（苹果）”的次序需要出现在“品类维度（手机）”之前。“是否为标识性维度”和“是否为时间维度”两个布尔变量非常关键，ETL 程序将根据取值为真的维度和原子指标字段做笛卡儿积，从而拼装成派生指标。

需要注意的是，该表一个必要配置的维度变量名为 updatetime，对应源数据集定义表中定义的增量抽取时间戳。

表 38 源数据集维度定义表

表名称	数据项名称	数据项类	详细定义	数据取值范围	主键

		型			
UIDS_DATASET_DIM 数据集维度定义表	数据集编码	String	源数据集的编码		Y
	维度变量名	String	原子指标的编码		Y
	维度类型	String	维度对应源数据集的字段名称，供 ETL 程序读取		
	维度次序	Int	维度在原子指标中的逻辑次序		
	维度展示名称	String	维度值在拼装为派生指标中的展示名称	CNY, USD, EUR, …	
	是否标识性维度	Bool	该维度是否为标识性维度	T:是标识性维度; F:非标识性维度	
	是否时间维度	Bool	该维度是否为时间维度	T:是时间维度; F:非时间维度	

10.4 ETL 作业步骤简述

根据通用指标配置模型，ETL 作业可批量读取源数据集的数据和原子指标定义，自动化地生成标准化的派生指标模型。整个流程可分为三大主要流程模块：

- 整理数据获取脚本并执行（下图黄色流程）
- 生成最细粒度派生指标即其时间序列数据（下图绿色流程）
- 整理指标描述类信息并写入存储介质（下图红色流程）

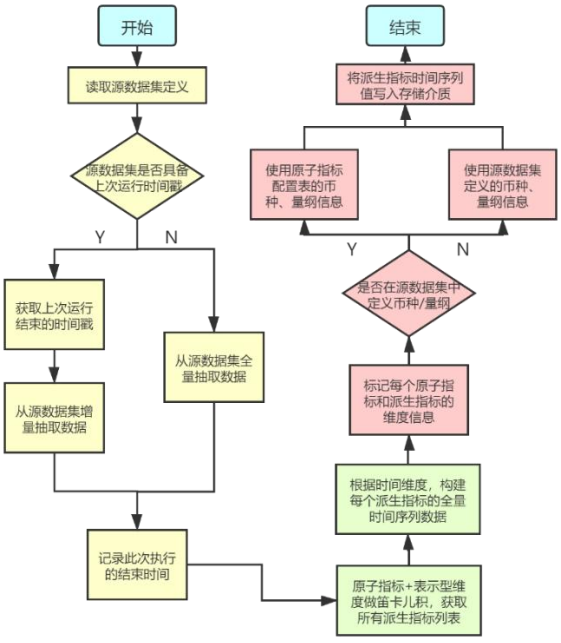


图 8 ETL 作业流程示意图

10.5 模型验证

考察通用指标配置模型符合实际业务的情况，包括对现有业务的覆盖情况和对未来业务的兼容性。以数个有代表性的源数据集为例，运行 ETL 数据作业，在生成的派生指标模型进行检验，检查该源数据集的业务含义在派生指标模型中的体现情况，和业务中涉及的数据维度、原子指标的整理提炼情况。

参 考 文 献

[1] 证监会发布《证券期货业数据模型第 1 部分：抽象模型设计方法》金融行业标准

[2] 关于征求《证券期货业数据模型 第 4 部分：基金公司逻辑模型》金融行业标准立项意见的通知

[3] ISO 3166-1:2006 Codes for the representation of names of countries and their subdivisions — Part 1: Country codes

[4] 国家统计局：统计用产品分类目录

[5] 中华人民共和国金融行业标准《上市公司分类与代码》JR/T 0020—2004