中国物品编码中心

一体化开发平台设计与建设规划总体标准及规范手册

V1.0

目录

[1. 建设目标 - 2 -](#_Toc121949482)

[2. 总体框架 - 2 -](#_Toc121949483)

[3. 总体标准 - 2 -](#_Toc121949484)

[3.1 标准规范体系 - 2 -](#_Toc121949485)

[3.2 业务术语 - 4 -](#_Toc121949486)

[4. 数据标准 - 5 -](#_Toc121949487)

[4.1 数据库设计规范 - 5 -](#_Toc121949488)

[4.1.1 数据库设计 - 5 -](#_Toc121949489)

[4.1.2 数据仓库设计 - 8 -](#_Toc121949490)

[4.1.3 数据库设计命名规范 - 13 -](#_Toc121949491)

[4.2 大数据应用规范 - 15 -](#_Toc121949492)

[4.2.1 分布式计算 - 15 -](#_Toc121949493)

[4.2.2 分布式存储 - 16 -](#_Toc121949494)

[4.2.3 数据采集 - 16 -](#_Toc121949495)

[4.2.4 数据加工 - 17 -](#_Toc121949496)

[4.2.5 数据挖掘 - 18 -](#_Toc121949497)

[4.2.6 数据可视化 - 19 -](#_Toc121949498)

[4.2.7 数据服务 - 20 -](#_Toc121949499)

[4.2.8 数据沙箱 - 20 -](#_Toc121949500)

[4.2.9 在线分析 - 20 -](#_Toc121949501)

[4.3 数据分类 - 21 -](#_Toc121949502)

[4.3.1 指标数据 - 21 -](#_Toc121949503)

[4.3.2 主数据 - 22 -](#_Toc121949504)

[4.3.3 通用基础数据 - 22 -](#_Toc121949505)

[4.4 数据资源编码规范 - 23 -](#_Toc121949506)

[4.4.1 指标数据编码 - 23 -](#_Toc121949507)

[4.4.2 主数据编码 - 23 -](#_Toc121949508)

[4.4.3 通用基础数据编码 - 23 -](#_Toc121949509)

[5. 技术标准 - 23 -](#_Toc121949510)

[5.1 技术架构规范 - 23 -](#_Toc121949511)

[5.1.1 开放性 - 23 -](#_Toc121949512)

[5.1.2 扩展性 - 25 -](#_Toc121949513)

[5.1.3 安全性 - 25 -](#_Toc121949514)

[5.1.4 可靠性 - 25 -](#_Toc121949515)

[5.2 性能设计规范 - 26 -](#_Toc121949516)

[5.2.1 性能指标 - 26 -](#_Toc121949517)

[5.2.2 可靠性指标 - 26 -](#_Toc121949518)

[6. 接口标准 - 28 -](#_Toc121949519)

[6.1 技术方式 - 28 -](#_Toc121949520)

[6.2 报文格式 - 28 -](#_Toc121949521)

[6.3 安全管理 - 29 -](#_Toc121949522)

[7. 应用框架规范 - 30 -](#_Toc121949523)

[8. UI设计风格规范 - 31 -](#_Toc121949524)

引 言

为深入贯彻中共中央、国务院《关于加强中国特色新型智库建设的意见》，全面落实集团公司“一个目标、三型五化、七个一流”总体发展战略，在《国家能源集团技术经济研究院建设一流智库行动计划（2021-2025年）》的总体指导下，按照集团公司《关于加强网络安全和信息化工作的指导意见》的工作要求，通过技术中台标准体系建设，为技经院智能化一流智库一体化平台建设与管理奠定基础，为后续其他课题建设提供指导。

# 建设目标

技术中台标准体系框架参考国标、行标、借鉴领先企业实践经验，并在融合优化集团已有标准体系成果的基础上，以构建“平台化、组件化、标准化”的业务应用为设计原则，承载智库平台整体的技术中台规范体系建设，包括数据规范、技术标准、接口标准、应用框架规范等中台相关技术规范。

# 总体框架

智能化一体化平台的标准体系主要分为总体标准、技术标准、应用框架规范三大类。信息化标准规范体系总体架构,如下图所示。

# 总体标准

总体标准是技经院智能化一流智库建设所需的总体性、通用性的标准和规范，应满足技经院智能化一流智库建设的总体设计、总体规划的要求，包括标准规范体系和业务术语。

## 标准规范体系

本次标准规范体系规定了技经院智能化一体智库标准规范框架及明细表。

|  |  |  |
| --- | --- | --- |
| 序号 | 规范标准分类 | 名称 |
| 1 | 数据库设计规范 | 应用系统数据库设计 |
| 2 | 数据仓库设计 |
| 3 | 数据库设计命名规范 |
| 4 | 大数据应用规范 | 分布式计算 |
| 5 | 分布式存储 |
| 6 | 数据采集 |
| 7 | 数据加工 |
| 8 | 数据挖掘 |
| 9 | 数据可视化 |
| 10 | 数据服务 |
| 11 | 数据沙箱 |
| 12 | 在线分析 |
| 13 | 数据分类 | 指标数据 |
| 14 | 主数据 |
| 15 | 通用基础数据 |
| 16 | 数据资源编码规范 | 指标数据编码 |
| 17 | 主数据编码 |
| 18 | 通用基础数据编码 |
| 19 | 技术标准 | 技术架构规范 |
| 20 | 性能设计规范 |
| 21 | 性能指标 |
| 22 | 可靠性指标 |
| 23 | 接口标准 | 技术方式 |
| 24 | 报文格式 |
| 25 | 安全管理 |
| 26 | 应用框架规范 | 应用技术分层架构 |
| 27 | 框架技术选型 |
| 28 | 总体应用要求 |
| 29 | UI设计风格规范 | 移动端设计规范 |
| 30 | PC端设计规范 |

## 业务术语

该类标准规定了技经院核心业务相关术语，包含选词原则、术语全称、简称、解释等。

* 智库：又称头脑企业、智囊集团或思想库、智囊机构、顾问班子，是指专门从事开发性研究的咨询研究机构。
* 研究项目：指通过中央和国务院相关部委及研究单位，以及集团公司、技经院或其他单位在科研计划或经费计划中安排，由技经院承担或参与，通过技术服务、技术咨询等合同（计划任务书）约定，并在规定时间内实施的软科学研究、技术开发、工业性试验、高技术产业化和成果应用等活动。
* 项目评估：通过集团公司相关业务职能部门委托，由技经院承担或参与，在规定时间内实施的评估咨询活动。
* 评估报告：评估报告是在尊重客观事实和科学规律基础上，有确切的评判依据和深入的论证分析，做到科学、客观、公正。
* 评估档案：评估档案包括项目委托书、可研报告及相关附件、专家意见、项目评审纪要、评估报告、项目完成反馈意见书以及其他相关资料。
* 评估专家费用：指在评估类项目评估过程中支付给临时聘用专家的咨询费。
* 情报服务：情报服务是情报机构为用户提供情报的工作，主要目的帮助用户克服搜集、选择过程中的困难，及时获取情报信息。
* 情报产品：是经过情报劳动加工而生产出来的产品，主要包括日报、周报、月报、季报、年报和专题专报等。

# 数据标准

数据标准制定是一体化开发平台建设的重要基础性工作，其目的是为编码中心核心业务发展提供统一的信息视图、数据规范及编码标准，实现跨应用、跨业务协同的信息共享，为战略研究、项目评价、信息服务三大主业提供支撑，助力编码中心向一体化快速迈进。

## 数据库设计规范

结合当前实践的研究，数据库设计规范分类分为面向应用系统的关系型数据库的数据和面向数据仓库的数据。关系型数据是指面向关系型数据库的数据，海量规模数据是面向数据仓库存储的数据。数据库标准+数据仓库标准可满足面向未来大数据应用、编码中心各类应用。

### 数据库设计

#### 设计原则

* 按阶段实施并形成项目阶段的成果物
* 符合3NF式要求，同时兼顾合规与效率
* 使用编码中心内规定的数据库设计软件工具
* 命名符合编码中心内的标准和规范

#### 设计过程规范

##### 数据分析阶段

在数据分析阶段，应注意搜集和分析数据相关的容，并形成相关成果物，包括数据流图和数据字典等，以此作为数据库设计的根底和依据。数据流图从数据传递和加工的角度，以图形的方式刻画数据流从输入到输出的移动变换过程。数据字典是对数据流图中的各种成分进展详细说明，作为数据流图的细节补充。数据字典一般应包括对数据项，数据结构、数据流、数据存储和处理过程的说明。

数据项描述={数据项名，数据项含义说明，别名，数据类型，长度，取值范围，取值含义，与其他数据项的逻辑关系}；

数据结构描述={数据结构名，含义说明，组成：{数据项或数据结构}}；

数据流描述={数据流名，说明，数据流来源，数据流去向，组成：{数据结构}，平均流量，高峰期流量}；

数据存储描述={数据存储名，说明，编号，流入的数据流，流出的数据流，组成：{数据结构}，数据量，存取方式}；

处理过程描述={处理过程名，说明，输入：{数据流}，输出：{数据流}，处理：{简要说明}}。

##### 概念设计阶段

* 目的

在数据分析的根底上，使用E-R模型技术，将现实世界中的客观对象抽象为实体和关系，形成概念数据模型〔CDM〕。CDM具有较强的语义表达能力，能够方便、直接地表达应用中的各种语义知识，应该简单、清晰、易于用户理解，是用户与数据库设计人员之间进行交流的语言。

* 方法

有自顶向下、自底向上、逐步扩张、混合策略的方法。

* 概念设计过程

集成

抽象

全局E-R模型

局部E-R模型

需求分析数据

第一步：数据抽象与局部E-R模型设计

1. E-R方法：建立E-R图；
2. 数据抽象：在多层数据流图中选择一个适当层次作为设计E-R图的出发点；确定每个局部应用包含哪些实体，实体包含哪些属性，实体之间的联系；划分实体和属性。

实体和属性划分的检验标准：

|  |  |  |  |
| --- | --- | --- | --- |
| 构成要素 | 检验基准 | 构成要素 | 检验基准 |
| 核心实体 | 集合性 | 核心关系 | 可选性 |
| 识别性 | 关系的度 |
| 永久性 | 业务指向性 |
| 使用性 | 核心属性 | 原子性 |
| 关系性 | 含义性 |

1. 局部E-R模型设计：

第二步：全局E-R模型设计

1. 集成各局部E-R模型，形成全局模型；
2. 集成的方法：多元集成发、二元集成法。

##### 逻辑设计阶段

* 目的

将E-R模型转换DBMS(数据库管理系统)支持的数据模型，逻辑模型要以数据为中心进行设计,并且必须在充分理解业务需求的基础上进行设计。设计的逻辑模型要具有扩展性强、一致性高、共享性强等特点。设计逻辑模型不仅仅是为了构建数据库，更重要的是起到对业务进行细致分析、准确定义的作用。

* 逻辑设计过程

逻辑模型整合

属性定义及范式化

实体选定及关系定义

从理论上来说，转换过程一般有7个步骤：

* 转换强实体
* 转换弱实体
* 转换1：1关系
* 转换1：N关系
* 转换M：N关系
* 转换多值属性〔Multi-Valued Attribute〕
* 转换n元关系〔n-ary Relation〕

E-R模型和关系模型的映射如下：

|  |  |
| --- | --- |
| **E-R模型** | **关系模型** |
| 实体类型 | 关系 |
| 1：1或1：N关系类型 | 外键 |
| M：N关系类型 | 两个外键 |
| n元关系类型 | n个外键 |
| 简单属性 | 属性 |
| 复合属性 | 简单属性的集合 |
| 多值属性 | 关系和外键 |
| 关键属性 | 主键/唯一键 |

##### 物理设计阶段

* 目的

基于给定的DBMS建立面向计算机物理表示的模型，描述了数据在储存介质上的组织结构，它不但与具体的DBMS有关，而且还与操作系统和硬件有关。

* 方法和过程

可以用建模工具直接将LDM转换为PDM。对于关系模型来说，进展物理数据库设计主要包括：

* 使用逻辑模型建立一系列的表〔如果在CDM和LDM中使用中文，应在转换后使用英文〕；
* 使用索引以提升性能；
* 实施约束和安全限制；
* 对数据进展分区和分布式处理等。

### 数据仓库设计

#### 基本分层结构

本项目数据仓库应从存储的内容可分为STAGE层（接口层）、DWD层（细节沉淀层）、DWA层（汇总衍生层）、DM层（数据应用层）。

STAGE层：生产系统数据源的直接拷贝，由数据采集过程对数据源进行直接抽取，在格式和数据定义上不作任何改变。提供业务系统数据文件的临时存储，数据稽核，数据质量保证，屏蔽对业务系统的干扰。

DWD层：数据仓库的细节数据层，对STAGE层数据进行沉淀，减少抽取的复杂性，将数据进行集中。系统将数据按业务域分类存放，跟STAGE层的粒度一致，属于分析的公共资源。

DWA层：数据仓库的衍生汇总数据层，该层通过对DWD层数据进行处理（轻度汇总、衍生），提高后续数据处理和访问性能，其特点是面向应用但不直接支持应用，将应用过程中的常用信息进行共同沉淀和处理，与DWD库共同构成企业级数据仓库。

DM层：是一个主题的、集成的、相对稳定的、反映历史变化的数据集合，主要用于面向业务分析应用场景的数据应用。

#### 数据仓库的开发方法及步骤

##### 数据仓库开发周期

数据仓库的开发应用具有其特有的、完整的生命周期，其周期可以分成:

* 数据仓库规划分析阶段
* 数据仓库设计实施阶段
* 数据仓库的使用维护阶段

这三个阶段是一个不断循环、完善、提高的过程。在一般情况下数据仓库系统不可能在一个循环过程中完成，而是经过多次循环开发，每次循环都会为系统增加新的功能，使数据仓库的应用得到新的提高。

##### 数据模型设计方法

数据模型设计采用自顶向下与自底向上相结合的方法进行设计。

以企业级数据模型为指导，结合生产系统相关规范，参考业界标杆SID、行业成熟模型，先建立概念模型，再在概念模型的基础上进行细化设计逻辑模型，物理实现时依据逻辑模型针对具体的分析需求和物理平台采取相应的优化策略。

* 3NF设计方法

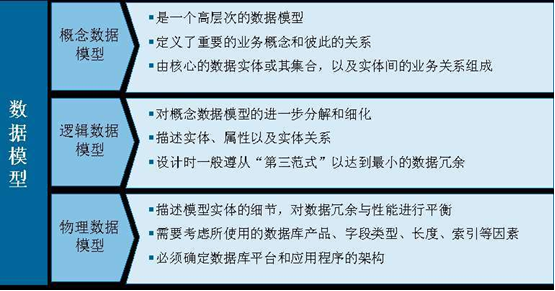
三范式建模是一种消除数据冗余的设计方法，数据被分成很多离散的实体，每一个实体在关系数据库中都对应一个数据表，可以通过实体来存储与这些特征有关的信息。

* 维度建模方法

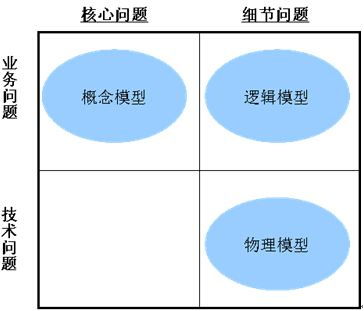
维度事实建模将客观世界划分为事实和维度，事实是由机构部门的业务过程和支持他们的业务源系统获得的，以数值形式出现并且具有可加性。事实由大量文本形式的上下文包围着，这些上下文被分割成多个独立的逻辑块，称其为维度。事实表由键和数值型度量（如收入、数量等）组成，因此具有健壮性和完整性的特点，维度表是用户分析数据的窗口，包含了事实数据表中事实记录的特性，有些特性是描述性信息，有些特性指定如何汇总事实数据表数据，以便为分析者提供有用的信息。

建模阶段

数据建模主要分为三个阶段：



不同的数据模型侧重解决不同层面的问题。



概念模型：概念模型是一个抽象的宏观层次（高层次、粗粒度）的业务模型，用于定义核心的业务概念实体和彼此的关系，最常用的是“实体-关系”图，所以在概念模型中最重要的对象是实体和关系。这一阶段主要工作是通过需求分析，明确需求所涵盖的业务范围，然后对需求范围内的业务及其间关系进行高度概括性的描述，把密切相关业务对象进行归类，划分为多个主题域。主题域是模型针对的业务的某一关注领域或关注点，同一个域内的实体具有高内聚性，不同域的实体之间具有低耦合性。

逻辑模型：逻辑模型是概念模型的延伸、分解和细化，表示概念之间的逻辑次序，是一个属于方法层次的模型。本阶段工作主要是根据业务定义、分类和规则对概念模型进行细化，定义其中的实体并描述实体之间的关系，产生实体关系图，然后遵照规范化思想在实体关系的基础上明确各个实体的属性。对于基础层一般采用3NF进行设计，达到数据冗余最小。对于汇总和信息子层一般采用反规范化冗余设计，快速支持数据访问和应用开发。

物理模型：主要依据逻辑模型针对具体的分析需求和物理平台采取相应的优化策略。此时会在一定程度上增加数据冗余或者隐藏实体之间的关系，是一种反规范化的处理，目的是提高数据分析的速度，适应具体数据库的容量、性能等限制。

##### 数据模型建模原则

为保证模型的稳定性和对业务支持的灵活性，建模阶段将遵循以下的原则：

继承性原则：以现有用户应用系统数据模型规范作为基础，在不影响理解情况下，尽量不提出新的概念；

稳定性原则：为保证模型的稳定性，实体与规则分离，突出核心实体的描述，提出规则点，对规则本身不做详尽描述；

前瞻性原则：为保证模型的前瞻性，同时采用自底向上和自顶向下的方式设计模型，其中自顶向下主要基于业务需求进行模型设计，使其可以完全覆盖到所有需求，自底向上主要基于业务逻辑而非业务需求进行模型设计，模型设计为囊括用户现有的各种业务关系，保证在有新需求时，底层模型能够对其进行支撑；

兼顾实际原则：在模型设计时针对当前用户系统建设的实际情况进行调研，保障模型能正常落地；

扩展性原则：模型的扩展包括实体属性的扩展和实体关系的扩展，模型设计为实体内只保留最细粒度的基本属性，粗粒度或上层的属性通过雪花结构的属性依赖关系实体来表现，这样在扩充属性或者扩充实体关系时，只增加表现属性依赖关系的实体即可。

##### 数据初始化策略

由于数据仓库的复杂性，针对不同的数据源采用不同的初始化策略。数据初始化后，系统运行一段时间，可能因为维护加载过程中的异常导致平台中数据与源端数据不一致，影响分析结果及对内决策的准确性。为保证核心数据准确性、一致性，应采取对账的手段保证数据加载过程的准确性，保证数据维护输入与输出端的一致性。

##### 数据资源存储策略

现有的数据源和目标数据种类繁多，有关系型数据库，文本文件、XML文件、视频文件、PDF/Office 文档等，可以大致归类为两种类型的数据，结构化的数据和非结构化的数据。

因此在数据存储和管理的设计中，可以利用数据库建立起结构化数据中心；根据非结构化数据的特点建立起非关系型数据库，对文档类信息进行存储。

1、关系数据库：包括OLTP数据库及OLAP数据库。OLTP（事务性）数据库主要用于存储结构化的关系型数据，支撑事务型应用，结构化资源数据存储在结构化数据数据库集群中。OLAP（分析型）业务采用MPP分析数据库，用于对现有的各领域数据从不同的纬度进行关联分析，用于实时查询。

2、分布式数据库：以高效访问为目的，没有实体之间的关联关系，侧重于可扩展性，大数据的读写能力。源系统的原始数据将直接存储在Hadoop数据平台的HDFS分布式文件系统中，用于对文档、图片等大量的文件进行分布式存储，在成本低廉的情况下能够保证存储的安全、吞吐率。。

##### 概念模型设计

概念模型是一个抽象的宏观层次（高层次、粗粒度）的业务模型，用于定义核心的业务概念实体和彼此的关系，最常用的是“实体-关系”图，所以在概念模型中最重要的对象是实体和关系。

概念模型设计常用的设计方法或者技巧如下：

* 抽象和继承设计方法：例如将A类重点人员、A类重点群体抽象出一个A类群体实体，将人员、群体的共同属性放在A类群体实体中，A类重点人员、A类重点群体独有属性放在自身的实体中。
* 多对多关系设计方法：例如客户和平台之间是多对多的关系，一个客户会在多个平台中办理业务，一个平台中有多个客户。

概念模型的最高层设计是主题域的划分，而主题域是模型针对的业务的某一关注领域或关注点，同一个域内的实体具有高内聚性，不同域的实体之间具有低耦合性。域的引入有助于建立模型框架的整体视图。

##### 逻辑模型设计

逻辑模型是概念模型的延伸、分解和细化，表示概念之间的逻辑次序，是一个属于方法层次的模型。具体来说，逻辑模型中一方面显示了实体、实体的属性和实体之间的关系，另一方面又将继承、实体关系中的引用等在实体的属性中进行展示。

逻辑模型设计如下：

* 对于DWD层采用3NF进行设计，达到数据冗余最小。
* 对于DWA层采用反规范化冗余设计，快速支持数据访问和应用开发。

在每个子域中首先明确本子域重点关注和解决的业务问题，然后细化子域的逻辑模型。随后对子域中的实体给出明确的定义，实体之间的关系给出解释，并且就实体的关键属性给出定义。

### 数据库设计命名规范

* 数据库文件名规范

命名方式：系统名\_文件名

* 用户

用户名的名称应该采用同系统应用相似的英文字符或字符缩写。从逻辑上分出的各数据层应当在数据库中用户实现其分离，规定如下：

* 数据库分区表规范

原则上数据仓库中的全量数据表都要建立分区（码表除外）。

* 数据库表索引规范

索引名命名方式：IDX\_表名\_字段名

* 数据库字段名规范

数据库字段命名按照对应英文单词大写英文字母定义，英文单词中间以下划线来隔开，如果英文单词较多则只取代表主要含义的单词，英文字母字数较多的一般截取前四位来简写。

* 数据库存储过程规范

（1）存储过程命名规则：“P\_”加上实体名称。

（2）存储过程要求有注释，注释内容为：列出创建人，创建用途，创建时间。

（3）存储过程日志规范：

每一存储过程均应记录执行存储过程的日志信息。必须调用专用写日志的存储过程，同时有exception时的处理机制。

（4）存储过程修改规范

修改时应注释清楚修改人，修改日期，修改原因和修改内容。

* 数据库函数命名规范

函数命名规则：F\_<功能名>

* 物理模型命名规范

（1）STAGE层

STAGE层接口表分各系统内部数据接口层和互联网爬虫数据两部分，两部分所有表的命名保持一致。

（2）DWD层

实体命名为前缀“DWD\_”加上“业务域代码\_分类代码\_内容描述”。

（3）DWA层

实体命名为前缀“DWA\_”加上“主题代码\_子类代码\_内容描述”。

（4）DM层

实体命名为前缀“DM\_”加上“应用代码\_内容描述”。

（4）临时表

测试或临时处理手动创建的表，为方便管理，统一命名规范

表名：T\_姓名缩写\_表标识[\_时间]

（5）中间过程表

为方便数据处理，临时创建的表

表名：数据分层前缀\_MID\_表标识

## 大数据应用规范

### 分布式计算

#### 实时计算

实时计算一般分为三个阶段：数据的产生与收集阶段（实时收集）、传输与分析处理阶段（实时处理）、存储对外提供服务阶段（实时查询）。实时计算根据需求分阶段底层支持以下能力：

* 支持配置；
* 支持容错；
* 支持管理工作进程和节点的故障；
* 支持水平扩展；
* 支持计算在多个线程、进程和服务器之间并行进行；
* 支持可靠的消息处理；
* 支持每个消息至少能得到一次完整处理，任务失败时，会负责从消息源重试消息；
*  支持低延迟、高效率消息处理；

#### 准实时计算

为支撑准实时应用，需提供准实时数据接入能力。保证数据的安全性、一致性和准确性。准实时计算根据需求分阶段底层支持如下能力：

* 高性能：能够以低资源消耗完成每秒数千交易的传送或者复制；
* 兼容性：适应各种异构大数据平台；
* 可靠性：保证数据的连续可用；
* 一致性：支持断点续传，恢复后自动从断点续传；
* 安全性：数据传输过程中采用压缩和加密；
* 高可用性：系统、机制、组织多维度保障业务高可用性，降低业务中断带来的损失。

#### 批处理计算

提供针对TB/PB级别数据、实时性要求不高的批量处理能力，主要应用于大型数据仓库、日志分析、数据挖掘、商业智能等领域。批处理应根据需求分阶段底层支持以下能力：

* 亿级数据 JOIN，作业 I/O 可达 PB 级/天；
* 具备跨集群数据共享能力，支持万级别的集群数，扩容不受限制；
* 提供功能强大易用的 SQL、M/R 引擎，兼容大部分标准 SQL 语法；
* 实现海量运算，用户不必关心数据规模增长带来的存储困难、运算时间延长等烦恼，智库平台根据用户的数据规模自动扩展机群的存储和计算能力，使用户专注于数据分析和挖掘，最大程度发挥数据的价值。

### 分布式存储

智库平台实现海量数据分布式存储，可存储结构化数据和非结构化数据。实现对数据存储的自动容错、集群存储在线扩容、系统数据访问安全控制，允许对数据节点进行热扩容，即在不停止分布式文件系统、不影响集群正常工作的情况下，进行节点的增加，提升集群存储能力。根据需求分阶段底层支持如下能力：

* 支持数据压缩；
* 提供大规模索引编制和搜索功能，可以构建分区索引、分布式索引和实时索引等，为各种分析尤其是文本分析的性能提供强大索引支持；
* 支持文件数据、流数据、互联网数据的分布式存储.;
* 支持在市场主流硬件平台上运行;
* 支持对底层分布式文件系统通过ACL提供安全性支持。;
* 支持管理节点单点故障热切换，保证集群稳定性。

### 数据采集

智库平台应提供支持数据实时或定时自动汇聚，支持通过动态规则配置采集技经院数据，实现数据的在线推送和实时汇聚。

#### 数据库采集

国家能源投资集团有限责任公司（以下简称“集团公司”）和国家能源集团技术经济研究院（以下简称“技经院”）内业务系统有大量的数据是存储在数据库中。智库平台根据用户定制的数据交换任务的需进行数据交换，从源数据库中进行抽取数据。

#### 流数据采集

智库平台在采集数据过程中会遇到以大文件形式存放的数据，以及从不同来源、不同服务应用持续不断到达且到达次序独立、数据来源众多格式复杂离散的数据流。这些文件及离散持续的数据需要送到不同的业务系统进行不同的加工组合。为采集和处理此类数据，平台采用Kafka消息中间件配合日志文件采集模块，再通过编程实现数据处理前置代理及流数据采集的业务处理解耦合。

#### 前置采集

智库平台从外部单位的业务系统采集数据时，应在数据源系统部署前置机和前置采集软件模块，实现前置机从数据源系统到智库平台的数据交换。

前置采集应提供以下功能：

* 支持多种数据源采集；
* 支持采集周期设置，按照分钟、小时和天设置不同的采集周期；
* 支持全量数据采集和增量数据采集两种采集方式。

### 数据加工

数据加工映射由转换器连接源和目标的数据流，实现完整的数据抽取、加工、载入的数据处理过程，基于可视化的数据流设计器，内置高度封装的映射组件，快速构建数据流的映射关系。根据需求分阶段底层支持如下能力：

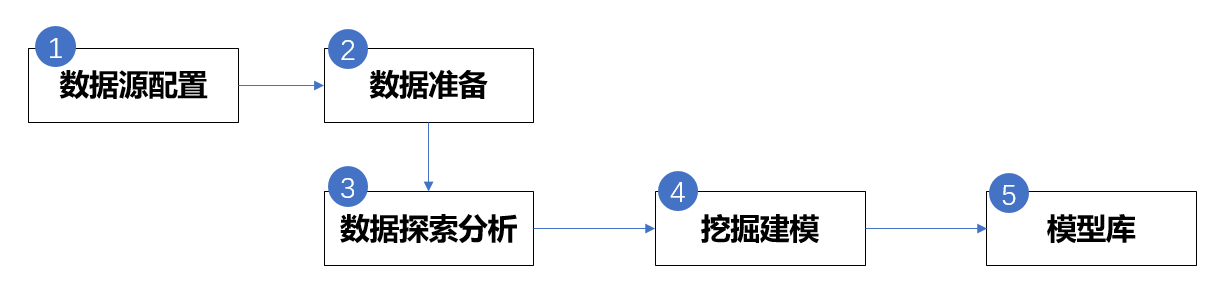
1、提供可视化的设计面板和映射组件，以鼠标拖拽的方式，方便快捷的编排数据流。

2、支持迁移映射实现不同存储方式之间的数据迁移，支持源、目标、筛选、表达式等组件，支持自定义映射可以自定义数据加工或处理脚本。

3、提供映射可视化管理，包括编辑、查看、签出、删除、样本测试、对象目录迁移。

### 数据挖掘

智库平台提供数据挖掘建模分析在线开发服务，实现数据挖掘模型的线上开发流程，开发完成的挖掘模型支持模型管理和外部调用，可视化数据挖掘根据需求分阶段支持如下能力。



可视化数据挖掘建模根据需求分阶段支持如下能力：

* 可视化建模。支持可视化方式拖拉拽组件完成建模，包括数据预处理、算法、模型评估等组件；
* 数据预处理能力。包括数据加载、数据选择、数据过滤、数据转换等功能，可在智能引擎界面将数据源连接数据选择后进入相关配置面板进行配置，将数据从一种表示形式变为另一种表现形式，例如：将连续变量转换为离散变量，将时间类型转换为时间差的数值类型或离散类型；
* 数据探索能力。采用类似EXCEL表格的方式管理和探索数据，包括对单变量的常用数学统计，根据数据字段选择后展示数据频次分布图；具备多维分析能力，对数据进行维度化分析后，完成多指标的散点矩阵图结果展示、多维度的分析、单维度多指标分析。实现特征分析、聚类分析、关联分析，情感分析，指数评估等支撑；
* 数据建模能力。支持数据建模的能力，分为训练模型和衡量模型进行预测两个过程。同时支持词库管理，主题采集等；
* 图像智能分析。利用数学模型并结合[图像处理](https://baike.baidu.com/item/%E5%9B%BE%E5%83%8F%E5%A4%84%E7%90%86" \t "_blank)的技术将图像中的目标从中分离出来，加以辨认、分析与追踪。如对文本资料图像进行分析处理，获取文字及版面信息；
* 算法能力。支持分类、聚类、关联分析、特征选择（如ReliefF算法）等算法，如朴素贝叶斯算法、随机梯度下降算法、K-means聚类算法、决策树、降维算法、特征抽取算法、线性回归算法、逻辑回归算法、随机森林算法等，并支持算法的分布式运算执行；
* 支持机器学习。平台的实时海量数据需依靠计算机批量处理，一定程度上赋予了计算机“学习”的能力。

### 数据可视化

数据可视化分析用于支持复杂的分析操作，可以根据分析人员的要求快速、灵活地进行大数据量的复杂查询处理，并且以一种直观而易懂的形式将查询结果提供给决策人员。数据可视化根据需求分阶段支持以下能力：

* 支持多维数据集选择

用户可以选择预先构建好的多维数据集。通过业务语义翻译，不需了解数据集存储的技术细节，即可以根据自己的业务需要及分析兴趣点选择多维数据集。选择过程中，可以按照多维数据集的名称进行过滤、快速定位。选择后，多维数据集进入工作区左侧的可用数据集区间。在可用数据集区间，数据集会进行结构展开，列出该多维数据集中包含的维度和度量。

* 支持多维数据集查询

一个多维数据集可能包含很多维度和度量，其中的维度和度量不一定都是用户所关心的。用户选择感兴趣的维度和度量，通过鼠标拖拽操作，将其放置于工作区表格中，即形成查询结果。

* 支持多维数据集钻取

用户发现数据现象后（如指标占比偏高、异常波动等等），需在更细粒度上进行分析，才能找到特殊数据表现的原因。系统支持用户选定数据行，然后选择维度进行下钻，在更细的数据颗粒度上观察以进行深入分析。在多维数据集的可用维度范围内，用户可以连续下钻，不断缩小数据范围，从而定位到对分析有帮助的最基础的数据单元。

* 支持多维数据集切片/切块

多维数据集中通常包含大量的数据，根据分析需要，在某些维度上限定维值，以便缩小数据集范围后，在其他维度上继续进行多维分析。

* 支持维度和度量定制

多维数据集中包含很多维度和度量，在运行时根据需求、中间分析结果来定制、切换维度和度量。选择不同的维度后，系统会根据选择的维度集合对数据重新进行汇总统计。

* 支持其他分析功能

通过公式输入产生计算列，考察指标的环比、同比，进行数据集旋转，钻透至关联数据集等。

* 支持多维数据集管理

多维数据集管理应提供将多维数据集的技术实现映射功能，帮助分析功能屏蔽技术细节，使分析功能仅需使用自己所熟悉的业务概念即可进行多维分析。对系统的数据集、维度、度量等进行统一管理，在维护业务语义的同时，确保数据的关联性及度量的口径一致性。

### 数据服务

智库平台通过构建标准化的数据开放服务，对外提供各类数据服务接口，实现对外数据服务的透明、可管、可控、可追溯。数据服务接口应包括：实时查询接口、批量同步接口、文件访问接口及消息订阅接口等。

### 数据沙箱

智库平台通过数据沙箱为工作人员提供一个可自由发挥的可靠的数据管理与运行环境，可进行数据挖掘、应用开发和程序测试。数据沙箱的管理应包括数据沙箱及状态的管理、内容组件的管理、数据的配置、测试环境、运行模拟管理等维度。

* 沙箱管理：管理数据沙箱的授权、回收，以及空间状态的管理；
* 内容组件：内容组件的使用及配置管理；
* 数据授权：数据范围、权限、数据的配置等管理；
* 测试环境：测试方法的配置等管理；
* 运行模拟：运行模拟工具及参数的管理。

### 在线分析

智库平台面向海量数据分析型应用领域，充分利用分布式存储和高效压缩技术，动态计算资源调配，提供高性能数据分析查询。这些数据通过分布在不同普通服务器上，被大量并发用户高速访问，满足技经院日益增大的海量数据分析需求。在线分析根据需求分阶段提供以下能力：

* 支持SQL-92标准；
* 支持PL/SQL标准存储过程；
* 支持基于Linux操作系统；
* 支持采用商用x86-64的标准PC服务器；
* 支持采用Master对等技术，系统高可用，弹性扩容；
* 支持海量数据的高效粗粒度索引机制，索引占用空间小，膨胀低；
* 支持分布式并行处理及内存计算；
* 支持PB规模以上的结构化数据存储；
* 支持高性能数据统计查询和数据加工处理；
* 支持高可扩展性。

## 数据分类

### 指标数据

1、指标数据定义

指标数据是组织在经营分析过程中衡量某一目标或事物的数据。指标数据通常是按照一定的规则和逻辑，对已有的原始数据进行加工和计算后形成的、具有统计分析意义的数据。

2、指标数据标准

指标数据标准确立集团公司层级统一使用的指标数据规范，是对指标的名称、业务含义、处理逻辑、数据格式等属性权威化、标准化、规范化的定义。

3、指标数据识别原则

搭建指标数据标准体系框架要按照找指标、理指标、用指标、管指标统筹考虑。其中：找指标是指找寻分散在企业各处的指标；理指标是指建立契合业务的指标体系；用指标是指规划指标体系在信息化中的应用方式；管指标是指建立常态化管理机制。

4、指标数据范围

* 指标数据范围涵盖产业：煤炭、火电、水电、风电、光伏、氢能、核能、铁路、港口、航运、化工、金融、装备制造和环保服务；
* 指标数据分为宏观国际、宏观国内、煤炭、石油、天然气、电力、可再生能源、集团汇报数据等。

### 主数据

1. 主数据定义

* 主数据是技经院需要跨系统、跨部门共享的核心业务实体数据；主数据具有属性相对稳定，准确度要求高，唯一识别等特点。

2、主数据标准

* 制定主数据标准是建立主数据代码库的基础工作，是实现系统之间数据共享的前提；是主数据管理组织及流程顺利开展的基石；
* 主数据包括员工主数据、组织架构等。

3、主数据识别原则

* 主数据特性：跨部门使用、生命周期长、跨系统使用、变动频率低、使用频率高；
* 共享程度：主数据通常会被多个业务系统访问，企业内使用同一主数据应用越多，越应确定是主数据；
* 业务价值：主数据描述企业最核心的业务，是企业最有价值的资产。

4、主数据范围

* 主数据范围包括课题研究、项目评价、情报服务三类技经院自有主数据；
* 主数据涵盖产业：煤炭、火电、水电、风电、光伏、氢能、核能、铁路、港口、航运、化工、金融、装备制造和环保服务。

### 通用基础数据

通用基础数据包括国家、地区、货币、计量单位等产业通用的数据及各产业特色基础数据。

## 数据资源编码规范

### 指标数据编码

指标数据编码是对指标数据的统一编号，本规范仅对技经院指标数据的编码规则进行约束。

指标数据编码由标识字符、分类编码两部分共9位字符组成，具体如下：

1. 一位标识字符：标识该指标数据的来源渠道。
2. 八位分类编码：
   1. 1位数据来源渠道；用英文字母缩写；
   2. 7位分类编码，编码用阿拉伯数。包括包括2位一级分类编码，2位二级分类编码，3位三级分类编码。

### 主数据编码

主数据编码由对应主数据归口管理部门进行主数据定义和指定编码规则管理。其中，编码规则包括编码规则、查重字段、编码业务解释等。

### 通用基础数据编码

通用基础数据在遵循《国家能源集团全产业数据标准体系框架规范（征求意见稿）》通用基础类数据标准体系框架基础上进行扩展编码。

# 技术标准

## 技术架构规范

技术架构应采用自主掌握的技术和国产化信创软件，遵循开放性、扩展性、安全性和可靠性原则，提升架构自主掌控和灵活配置能力。

### 开放性

开放性原则主要是从业务、技术、管理等方面充分考虑技术架构的开放性规范。

#### 业务开放性

在系统开放性设计中，为了最大限度地增强系统的价值，最大限度地吻合各业务应用的需求，通过对业务深入理解，对业务的发展方向进行预判，整个系统设计过程始终遵循面向业务价值，围绕系统应用，依靠业务部门，注重实效的方针，保证业务的开放性，满足业务需求不断发展变化的要求，便于应用系统的升级。

#### 技术开放性

技术开放性需要从组件化结构、标准化结构、开放功能包、分层架构和平台通用性方面设计

* 组件化结构

采用全组件化结构设计，每个组件都被独立地实现，并通过标准接口联系在一起。每个功能组件在功能上独立，同时可根据用户需求灵活配置、组合，实现平滑升级扩容，功能实体可使业务和开发人员根据具体使用要求增加或减少系统应用模块。

* 标准化接口

采用标准统一的接口设计，所有功能实体间的数据交换以及对其他模块的数据引用都通过标准接口完成，使多个组件对接时在开放性、稳定性、扩展性与集成性上有着很好的适配空间。

* 开放功能包

系统处理组件化结构设计与标准化接口设计以支撑开放体系结构外，为了方便用户个性应用的开发，还封装系统及其组件所需的二次开发应用工具包，使其他技术团队对平台进行二次开发时能够更好地复用。

* 分层架构设计

采用横向分层和纵向分割架构设计。将层与层之间相互分离，每层的应用和服务，采用独立的模块开发和部署，模块间交互标准化，新增功能模块分解到各层，以插件形式加入原系统，既不影响整体架构，也不影响本层功能提供，具备高模块化设计，保证了系统功能的可扩展性。纵向分割是将业务和可复用服务分离出来，通过分布式服务框架调用。新增产品可以通过调用可复用的服务实现自身的业务逻辑。

* 平台通用性

系统采用的软件开发技术可运行于通用的主流硬件平台上，不依赖于特定的、专用的硬件设备或者系统软件。系统配置（硬件系统、操作系统、数据库系统）的升级一般情况下，不会引起系统的修改和再次开发。

#### 管理开放性

管理开放性是指软件系统在业务流程、规则、审核和组织架构方面采用结构化配置和管理模式，根据系统管理流程，开放软件系统管理功能的可配置性，从而灵活快捷的支持管理流程开放性能力，适应不同的管理模式和流程。

### 扩展性

为了满足本项目的长远目标，适应业务和技术不断发展得需要，系统软件架构采用模块化体系结构，新功能、新业务的增加能够在不影响系统运行的情况下实现，使架构在一定范围内支持业务的快速变化，并能够适应新技术的应用。

随着用户和数据量增加，系统可能遇到资源瓶颈，如数据库出现容量和内存使用居高，应用服务器出现并发请求排队等。此时，可以通过横向扩展或物理分服机制提升系统的负荷能力。

* 横向扩展：通过集群的负载均衡机制，通过增加集群节点来传播负载和增加可靠性。作为系统扩展的主要方式，这将在绝大多数情况下满足系统的扩容需求；
* 物理分服：当用户和数据量持续增加到横向扩展无法提升的情况时，可以考虑通过物理分服的方式扩展系统。在系统总入口处通过用户路由到相应的服务群。

### 安全性

在设备选型和系统设计过程中，从安全架构、安全策略及安全机制等多方面充分地考虑安全问题。

系统及运行环境采取全面的安全保护措施，具有防病毒感染、防黑客攻击措施，同时在防雷击、过载、断电和人为破坏方面进行加强，具有高度的安全性和保密性。对接入系统的设备和用户，进行严格的接入认证，以保证接入的安全性。同时系统支持对关键设备、关键数据、关键程序模块采取备份、冗余措施，有较强的容错和系统恢复能力，确保系统长期正常运行。

### 可靠性

系统采用故障检查、告警和处理机制具有强的容错能力和错误恢复的能力，保证数据不因意外情况丢失或损坏，保证系统安全稳定可靠运行。

* 当操作错误时，能迅速有效地实现回滚。对系统硬件或网络故障，数据库系统也应能提供一定的保护措施，保证系统安全稳定可靠运行；
* 系统采用灵活的服务部署方案实现负载均衡，防止“瓶颈”产生；
* 系统具备多种冗余、备份和集群处理的机制和功能，关键部件、数据库具备冗余备份和负载分担机制，系统具备冗余配置，保证系统无单一故障点，且应易于扩容和维护；
* 系统提供完善的日志记录能力，对系统关键数据的每一次增加、修改和删除都能记录相应的修改时间、操作人和修改前的数据记录；
* 系统具有监控功能，能监视系统各功能模块的运行情况，随时发现系统自身的问题；
* 本系统提供容错机制，避免单点故障的发生，如采用集群、负载均衡等技术。

## 性能设计规范

### 性能指标

性能指标要求主要包括以下几个方面：

* 数据采集性能要求(在网络环境具备情况下)

结构化数据采集每秒大于10000条；

非结构化数据采集每秒大于一个80M文件。

* 页面调用性能要求(在网络环境具备情况下)

网内访问静态页面响应时间小于2秒，动态页面响应时间小于3秒；

网外访问静态页面响应时间小于3秒，动态页面响应时间小于5秒。

* 接口性能要求

并发处理：智库平台即可并发处理能力需要具备不低于100的并发。

### 可靠性指标

可靠性指运维团队快速响应7\*24小时服务能力，包括系统可靠性、接口可靠性、数据可靠性。

#### 系统可靠性

1、系统的Web服务器出现故障应能及时告警，应具有完整的操作权限管理功能和完善的系统安全响应机制。

2、 Web服务器应采用双机或者集群方式对外提供服务，在某台服务器出现故障时能保证页面能被正常访问。

3、通过前置负载均衡器，以一定的算法将外部请求分发到WEB应用，从而提升WEB服务的整体处理能力。

4、Web服务器升级时应通过负载均衡器在前端逐个隔离待升级节点，在业务节点升级完成后再将其加入集群。这样就能保证整个升级过程较为平稳，不会造成大量用户在升级期间无法访问业务。

5、为了减少异常断电对系统造成过大影响，尽量将Web服务注册到操作系统的自启动脚本中，从而保证操作系统正常运行后能自动启动应用，以便系统能尽快向外提供服务。

6、Web应用端口隔离，当不同的Web应用需要部署在同一主机时，应该尽量使每个Web应用享有独立的Web容器，达到各Web应用的进程级隔离目的；从主机上隔离会达到更好的资源隔离效果。

#### 接口可靠性

1、接口出现故障应能及时告警，应具有完整的操作权限管理功能和完善的系统安全机制。接口失败时应建立重试机制，出现故障时需要及时地响应。

2、接口升级、割接等过程中，应保证有系统回退机制。

3、接口应防止消耗过多的系统资源而使系统崩溃。

#### 数据可靠性

1. 数据完整性：要求智库平台提供的数据符合实体完整性约束、参照完整性约束。

2、数据及时性：要求智库平台中提供的内外部数据应是最新状态的数据。

3、数据一致性：要求智库平台提供的数据对内对外应保持一致。。

4、 数据准确性：要求智库平台提供的数据是准确的。

# 接口标准

## 接口要求

* 信息通讯安全

保证接口的自身安全，通过接口实现技术上的安全控制，做到对安全事件的“可知、可控、可预测"，是实现系统安全的一个重要基础。可对其进行：安全评估、访问控制、防恶意代码、加密等方法确保接口通讯安全。

* 支持高开发

系统在设计时考虑到多线程的情况，可采用了分布式架构和负载均衡技术，能够支持大规模的并发访问。

* 可监控

具备完善的监控功能，可以实时监控系统运行状态、用户访问情况等，及时发现并解决问题，保证系统的稳定性和可靠性。

* 系统资源的动态扩展

保证在充分利用系统资源的前提下，实现系统平滑的移植和扩展，同时在系统并发增加时提供系统资源的动态扩展，以保证系统的稳定性。

* 异常处理机制

表单验证、唯一性检查、或其他可预期的错误。我们需要编写特定代码来捕获这类错误，并抛出一个包含提示信息的全局异常，捕获并返回给客户端。

* 业务扩展

系统在运行过程中，根据需要扩展业务功能例如，当用户需求发生变化时，可以增加新的业务功能以满足用户需求。这种方式可以提高系统的灵活性和可扩展性，但是需要考虑业务变化的频率和影响范围。。

## 技术方式

提供的数据接口技术主要包括以下类型:

* HTTP接口

通过HTTP协议传输的接口，可以传输文本表单数据，也可以传输json类型的对象数据或xm1类型的数据。

* SOAP

简单面向对象协议，基于HTTP，使用xml作为默认传输格式。

* Web Serveice

相比传统的HTTP接口只传输文本请求和文本响应，通过Web Service可以直接拿到远程的一个对象，并能够直接调用对该对象的属性和方法，比HTTP更高级。

* FTP

FTP是用于在网络上进行文件传输的一套标准协议，客户在和服务器建立连接前要经过一个“三次握手”的过程，保证客户与服务器之间的连接是可靠的，而且是面向连接，为数据传输提供可靠保证。

## 报文格式

数据接口的报文格式主要分为报文内容、异常代码标准、报文格式说明规范组成。

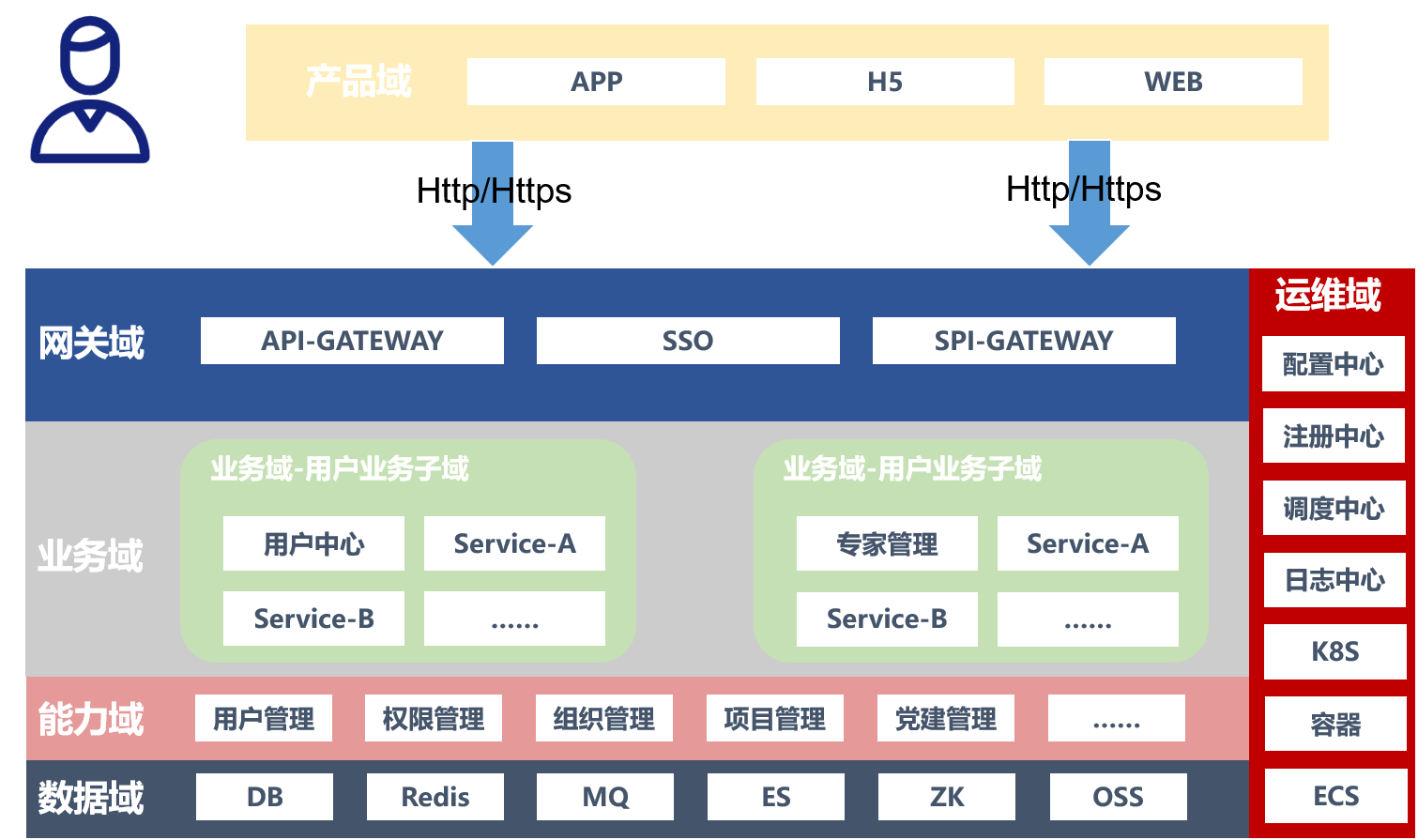
* 接口内容主要分为XML和JSON两种报文格式。
* XML可扩展标记语言，标准通用标记语言的子集，是一种用于标记电子文件使其具有结构性的标记语言。
* JSON(JavaScript Object Notation) 是一种轻量级的数据交换格式。大数据基础平台对外开放接口采用REST风格，使用HTTP+JSON报文方式承载交易信息，支持GET、POST两种请求方式。
* 接口异常代码标准

即调用数据服务接口异常场景下返回的错误码，返回异常码的同时也会返回中文详细描述，以增强异常码的可读性。异常码通常表示某种异常场景，具有唯一指向性。

* 报文格式说明规范
* 简介
* 接口描述
* 请求参数、响应
* 应用示例
* 异常码说明

# 应用框架规范

应用系统建设模式基于插件化和服务化，采用主流的轻量架构，可对服务和资源的整体管理。平台在PaaS层采用更加轻量的Docker容器，组件化架构采用微服务架构。



1、**⽹关层：**为业务平台提供内外网隔离、统⼀鉴权、限流、降级等，提高系统整体安全性、稳定性。

* 网关域仅负责通用平台的进出口请求业务分发、协议转换、权限控制等通用功能，不提供任何实际业务代码；
* 平台各个业务系统均不提供外网权限，外部系统主动请求平台通过API网关转发；
* SSO服务仅提供多渠道登录鉴权、分布式session、AccessToken验证功能，不提供其他业务逻辑。

2、**业务层：**基于能力层功能，完成平台各个业务流程组织，根据业务范围做业务编排。

* 业务域仅负责不同业务流程的组织，而不负责核⼼领域数据的持久化及状态维护（例如业务域负责微信授权注册、网站⼿机注册2种不同流程主旨，但是实际的⽤户创建和持久化通过调用能力层完成）；
* 业务域仅负责不同业务流程的组织，而不负责访问权限、登录会话等信息维护。（权限和会话统⼀由网关层统⼀维护）。

3、**能力层：**核⼼业务领域模型、标准作业流程维护，对外输出通⽤能⼒，⼀边个性化业务流程组织。

* 能力域各业务系统仅负责核⼼领域模型等持久化、状态流转（例如：⽤户服务）；
* 业务域各个系统通过RPC接⼝调⽤能⼒域的通⽤能⼒。（例如注册）。

4、**数据域层：**各个业务板块涉及的数据保存、缓存、等功能架构。

* 数据域统⼀由运维、数仓维护，各业务系统不再搭建数据存储服务；
* 各个业务系统根据业务需求申请特定规模的数据存储配额和账号；
* 数据域暂时以私有云服务为主，以提升系统底层数据稳定性。

5、**运维域：**平台运维管控相关业务架构，包括系统监控、运维保障等。

* 需提供运维相关⼯具类服务，不实际涉及业务逻辑。

# UI设计风格规范

详见《UI设计风格规范及原则分册》