数据集成开发框架

应用标准规范

第1部分

总体框架

北京神舟航天软件技术有限公司

二〇一八年八月

目录

[1 范围 1](#_Toc72849849)

[2 规范性引用文件 1](#_Toc72849850)

[3 术语和定义 1](#_Toc72849851)

[4 总体架构 1](#_Toc72849852)

[4.1 功能架构 1](#_Toc72849853)

[4.2 数据架构 1](#_Toc72849854)

[4.3 网络架构 1](#_Toc72849855)

[4.3.1 数据分析网 1](#_Toc72849856)

[4.3.2 专网 1](#_Toc72849857)

[4.4 接口架构 1](#_Toc72849858)

[5 技术要求 2](#_Toc72849859)

[5.1 开发技术选型要求 2](#_Toc72849860)

[5.1.1 开发语言要求 2](#_Toc72849861)

[5.1.2 统一系统软件要求 3](#_Toc72849862)

[5.1.3 开发体系要求 3](#_Toc72849863)

[5.1.4 安全可控要求 4](#_Toc72849864)

[5.1.5 IPv6支持要求 4](#_Toc72849865)

[5.2 统一认证及授权支撑服务要求 4](#_Toc72849866)

[5.2.1 统一认证服务要求 4](#_Toc72849867)

[5.2.2 分级管理员服务要求 7](#_Toc72849868)

[5.2.3 统一组织结构服务 8](#_Toc72849869)

[5.3 系统集成模式要求 9](#_Toc72849870)

[5.3.1 系统间服务集成模式要求 9](#_Toc72849871)

[5.3.2 系统间数据集成模式要求 11](#_Toc72849872)

[5.3.3 域名规划要求 13](#_Toc72849873)

[6 应用边界规则 15](#_Toc72849874)

[6.1 数字化指挥相关业务边界 15](#_Toc72849875)

[6.2 管理相关业务边界 16](#_Toc72849876)

[6.3 知识管理相关业务边界 16](#_Toc72849877)

[6.4 互联网信息采集相关业务边界 16](#_Toc72849878)

[6.5 身份认证管理相关业务边界 17](#_Toc72849879)

[6.6 统一运维管理相关业务边界 17](#_Toc72849880)

# 范围

本标准规范规定了工程三期项目中全部应用系统开发中需要遵循的总体架构约束，明确应用系统开发的技术路线选型策略，规定系统间集成接口标准规范等相关内容。

# 规范性引用文件

略

# 术语和定义

略

# 技术要求

为降低项目后续运维风险，本项目的技术路线规划采用的统一规划设计综合支撑服务平台提供共性支撑服务，规范化应用开发过程技术选型方案降低运维和二次开发技术门槛，统一规定基础软件使用范畴满足统一规划要求。

## 开发技术选型要求

本项目中大数据中心、管理系统、现场系统、指挥中心系统和模拟仿真系统的开发体系存在较大差异。例如：大数据中心的技术体系为数据处理和数据集成类应用，管理系统为基于表单、流程的管理信息系统，现场系统为基于终端的CS应用系统，指挥系统为基于大屏的可视化系统等。

为充分发挥各开发厂商自有的开发技术优势，本规范中不对具体的应用级平台进行强制性约束，仅对开发中的基础开发技术选型提出要求，目标在于降低后续持续运维工作技术栈复杂度，同时也保证项目进行IPV6改造、跨平台移植、国产化改造等相关工程建设要求时，避免产生颠覆性架构调整情况。

### 开发语言要求

本项目开发可使用开发语言仅可使用如下开发语言：Java8、JavaScript、HTML5、CSS、C语言、SQL，所有Java程序运行的JDK版本为Java SE 8u181。

### 统一系统软件要求

|  |  |  |  |
| --- | --- | --- | --- |
| 编号 | 系统软件类型 | 系统软件名称 | 系统软件版本 |
| 1 | 事务型数据库 | 神通数据库管理系统 | V7.0企业版 |
| 2 | 分析型数据库 | 神通数据库管理系统（MPP集群版） | V7.0 |
| 3 | 桌面端操作系统 | Windows操作系统、LINUX操作系统 | Windows7 32位、Windows7 64位、Window Server 2012 R2 64位（虚拟应用）、Windows10 32位中国政府版、Windows10 64位中国政府版、中标麒麟桌面版 |
| 4 | 服务端操作系统 | Windows操作系统、LINUX操作系统 | Window Server 2012 R2 64位、中标麒麟高级服务器操作系统软件V7.0 |
| 5 | 浏览器 | 专用安全浏览器（海泰红莲花浏览器安全定制版） | Chrome 64 |
| 6 | 应用中间件 | 东方通应用服务器软件（TongWeb） | V6.1 |
| 7 | 消息中间件 | 东方通消息中间件软件（tonglink/Q） | V8.1 |
| 8 | 云平台 | 华为FusionSphere | V6.1 |
| 9 | 电子签章 | 金格iSignature电子签章系统 | V8.2 |
| 10 | 杀毒软件 | 360网神终端安全管理系统 | V7.0 |
| 11 | 地理信息库 | 超图云GIS应用服务器平台[简称：SuperMap iServer] | V9 |
| 12 | OFD版式文件转换工具 | 数科网维OFD版式文件转换工具 | 略 |
| 13 | 流式文档编辑工具 | 金山WPS Office | 2016专用版 |
| 14 | 统一融合通信工具 | 华为CloudEC | 6.1.0 |

### 开发体系要求

|  |  |  |  |
| --- | --- | --- | --- |
| 编号 | 应用表现形态 | 架构类型 | 架构选型 |
| 1 | CS应用 | 客户端界面 | JAVA SWT/JAVA SWING |
| 2 | WEB应用 | Web服务端框架 | Spring MVC/Spring boot |
| 3 | 应用服务 | 微服务框架 | 华为ServiceComb 1.1.0 |

### 安全可控要求

本项目中相关应用系统开发中服务端程序和终端程序均须支持安全可控要求，能够支持安全可控CPU（包括：飞腾、龙芯、申威等）、安全可控操作系统（包括：中标麒麟等）。

### IPv6支持要求

本项目中相关应用系统开发中服务端程序和终端程序均须支持IPv6协议，能够满足IPv4和IPv6双栈运行能力。初期部署基于IPv4进行实施，应用系统单项初验前需要完成IPv6改造升级。

## 统一认证及授权支撑服务要求

### 统一认证服务要求

本项目中相关应用系统必须同统一认证服务集成，实现对用户业务身份的统一认证。综合服务支撑系统中统一认证与授权管理分系统提供统一身份认证服务，支持与现有CA认证系统的数字证书认证进行集成，并支持用户名密码等多种认证方式，支持单点登录、登出。

#### 统一认证过程

实现统一认证/SSO服务后的部署方式如下图所示。

1）客户端：客户端包括专用的安全浏览器、门户桌面客户端或者其他客户端应用；

2）安全网关：由CA提供的安全网关，访问应用系统可能通过安全网关，也可能不通过，两种情况下访问时携带的请求信息是不同的；

3）应用系统：多个需要接入SSO的应用系统，相当于SSO Client；

4）SSO Server：提供统一口令认证服务，认证成功后，发放令牌。

#### 单点登录过程

统一认证服务（SSO Server）实现身份认证并生成认证结果令牌（sso\_token），令牌将交由客户端（安全浏览器其他客户端）保存。随后对所有应用系统的访问均应携带此令牌，应用系统识别并验证此令牌中携带的认证结果信息即可。基于安全考虑，统一认证服务需要保证令牌不可伪造。

一个典型的单点登录过程如下图所示：

单点登录机制示意图

1）客户端：即用户代理，比如安全浏览器、门户桌面客户端等。客户端需要在未用户未登录的情况下，引导用户去SSO Server做口令认证，并在客户端Cookie中保存认证令牌sso\_token，之后，请求中将自动携带sso\_token。

2）SSO Server：提供统一口令认证服务，认证成功后，发放令牌。

sso\_token={payload:base64(用户信息),signature:私钥加密}

3）SSO Client：应用系统不再需要登录页面，而是处理请求前，识别请求中有效sso\_token里的用户信息,否则让客户端重定向去做认证或拒绝请求。为了防止sso\_token伪造，SSO Client需要对sso\_token的signature使用公钥解密。如果能解开，证明该sso\_token是由SSO Server发放的，是合法的。

基于上述典型过程，单点登录流程如下图：

上述单点登录过程中，客户端获取令牌后，访问其他应用系统时必须携带令牌，否则服务端将无法识别请求，出于安全性考虑通常会拒绝访问。

#### 各类客户端访问要求

根据客户端访问应用系统方式的不同，携带令牌的方式至少包括下列几种：

1）安全浏览器直接访问：直接通过安全浏览器访问多个应用系统时，如果这些应用系统具有统一的一级域名，则安全浏览器可以自动携带令牌访问，实现单点登录；因此：在直接通过浏览器访问应用系统时，要求接入SSO的应用系统配置与SSO Server统一的一级域名（二级域名可以不同）；

2）客户端正向代理：如果使用CA提供的客户端正向代理功能，则无论是否通过域名访问应用系统，都会携带CA认证完成后的令牌，接入SSO的应用系统只需要能够识别这个令牌即可。这也是目前已经与CA集成的应用系统所采用的实现方式，本文实现的SSO方案继续兼容；

3）门户桌面客户端和其他客户端：此类客户端在获取令牌后，继续访问其使用的应用系统时都应该携带令牌，便于应用系统服务端进行身份认证。

统一认证服务的详细说明及接口要求，请参照《统一认证及权限管理开发接口规范》

### 统一组织结构服务

组织结构管理可以按照署的组织架构，统一管理所有机构，并支持以组织机构树的形式展现。各级组织在组织结构树中以最小行政粒度为叶子节点，例如：署本级组织机构到处级。署本级统一组织结构服务中规定仅包括署本级人员（包括内设机构、驻地方特派员办事处、直属单位、派出局等），如署本级统一组织结构中需要加入地方机关人员（用户），需在署本级统一组织结构树中创建地方机关组织机构树节点，并将人员加入到对应的组织结构下，下图给出组织结构管理和用户管理两个界面示意图。

一个人可关联多个组织，但每个人仅有一个在编组织，查询指定人员（用户）的组织时将返回该人员的在编组织和其他关联的组织。

人员（用户）属性中包含职级属性，职级信息通过数据字典定义，如司级、副司级、处级、副处级、科级、副科级等。

三期信息系统实现署、省两级部署，在地方上的人员若同时需要使用署的系统和省级厅的系统，则该人员将分别有署、省两级系统签发UKey，且该人员的两个UKey不能混用。

在统一权限管理系统中，当新增、删除、编辑组织和用户信息时，将通过消息中间件向各应用系统发送消息，各应用系统需根据管理需要进行数据同步处理。消息格式请参照《统一认证及权限管理开发接口规范》。

## 系统集成模式要求

### 系统间服务集成模式要求

本项目中各系统间的服务集成的主要模式包括：域内应用服务间集成模式、业务访问服务集成模式、跨系统应用页面集成模式、跨网闸应用服务间集成模式、应用间异步消息集成模式和成品软件专用API集成模式。

#### 域内应用服务间集成模式

域内应用服务间集成模式主要应用于可通过IP地址直接双向连通的同一网络域范围内各应用系统后端服务间的集成成访问要求。技术路线上采用基于华为云提供的PAAS层微服务平台ServiceComb支撑能力，满足服务注册、集中配置、容错限流、服务认证、服务监控等应用间服务集成基础支撑架构要求。微服务开发采用Spring Boot标准兼容模式，降低微服务开发学习门槛，通过修改配置方式无缝接入ServiceComb运行体系。

#### 业务访问服务集成模式

业务访问服务集成模式主要应用于用户请求发起端（包括：应用系统前端Web页面、CS客户端程序和Web服务器端程序等）访问公共应用业务服务所使用的集成模式。针对此类服务集成采用微服务网关方案，请求发起端在发送请求时需要将原发起端获得的认证令牌信息发送给微服务网关，并在微服务网关层实现对业务访问用户的统一身份认证和服务级统一授权访问控制。

#### 跨系统应用页面集成模式

跨系统应用页面集成模式主要应用于跨应用系统间的Web页面共享访问集成，例如：数字化指挥系统直接调用业务管理系统的项目资料查看功能页面。针对此类服务集成采用带令牌身份的请求拦截技术方案，请求发起方在请求的Web地址上拼接令牌参数，并由统一授权系统提供的拦截器进行统一拦截，并进行身份认证，整个过程对被请求端完成透明，以避免被请求端重复开发工作。

#### 应用间异步消息集成模式

应用间异步消息集成模式主要用于应用系统间在同一机关内或跨机关进行服务异步通信集成。技术层面采用在多级机关统一部署东方通tonglink/Q分布式消息中间件，统一应用间异步消息集成要求。例如：1、署本级组织机构变化时可通过此服务向各系统发送通知消息；2、行政办公系统通过此服务完成署级机关向省级机关下发公文信息等。

#### 跨网闸应用服务间集成模式

跨网闸应用服务间集成模式主要应用于跨等保安全域间的服务请求。在技术路线上采用基于网闸文件交换服务，有应用服务将服务请求转换为文件输出到网闸交换目录，并通过网闸交换策略，定时（秒级触发）将文件交换到另一等保安全域的网闸交换目录，并由跨网闸应用服务扫描交换目录中的新增请求文件，解析并处理服务请求，将处理结果上传到网闸交换目录，网闸文件交换服务将结果文件通过网闸交换策略，定时（秒级触发）将结果文件交换到原等保安全域的网闸交换目录，原请求服务扫描交换目录中的新增结果文件，解析并返回服务请求结果。

上述服务集成过程由各应用系统根据实际使用需求自行设计请求文件和结果文件定义格式，交换目录文件扫描程序和处理转发程序由各应用系统自行实现。

网闸交换两端所部署的跨网闸应用服务需向总包方备案并通过安全审查后方可部署。

#### 成品软件专用API集成模式

成品软件专用API集成模式用于同本项目采购的华为eSpace统一融合通信软件、华为云平台软件、神通数据库软件、数据挖掘软件等成品软件间集成，应用软件须根据成品软件提供的API接口规范，实现应用同成品软件间集成。

### 系统间数据集成模式要求

各系统间数据集成的主要模式包括：非结构化数据交换集成模式和数据ETL处理集成模式。

#### 非结构化数据交换集成模式

非结构化数据交换集成模式主要用于被单位到专网交换区、专网内部服务器间、专网交换区到数据分析网交换区间、数据分析网数据报送终端到数据分析网交换服务器、数据分析网服务器间、标准两级交换中心服务器间的非结构化文件交换需求。能够支持按预先规划的数据路由完成交换调度，也能够支持通过传输代理组件主动推送数据模式。在传输文件大小方面能够支持小文件和TB级大文件交换，在网络模式方面能够支持局域网、广域网、专线接入网络、电子政务外网等多种网络环境。

非结构化数据交换集成模式针对署省两级交换中心服务器间的非结构化文件交换需求，如无大文件高速交换要求，可使用东方通TongDXP数据交换平台。

针对其他大文件传输场景，在专线接入网络环境下，目前尚未找到合适的解决方案。目前，正在调研友友DataEcho数据交换平台在此需求场景下的满足程度。另外还可以考虑的解决方案包括：1、应用层自行完成上传文件本地切割，并基于单线程数据交换中间件实现切割后文件的分片上传，在服务端完成切片文件合并，并完成合并后文件的完整性校验，需要应用层做较多开发，对系统开发周期有较大影响；2、自行开发单文件多线程、多通道数据传输中间件，组织独立团队进行技术攻关，短期内达到一定稳定程度具有一定风险，底层通讯层基础软件开发人员相对匮乏，团队短期内组建有一定困难。

#### 数据ETL处理集成模式

数据ETL处理集成模式主要用于同一网络域内进行数据清洗转换集成使用，可利用数据ETL处理集成模式实现标准库构建、主题库构建和指挥系统数据抽取集成等业务目标。

# 数据应用标准

## 分类

分类的目的是通过建立分类模型(称作分类器)，把数据库中的数据项（记录）映射到给定类别中。要构造分类器，需要有一个训练样本数据集作为输入。由于数据挖掘是从原数据集中挖掘知识的过程，这种类知识也必须来自于原数据，应该是对原数据的过滤、抽取(抽样)、压缩以及概念提取等。从机器学习的观点，分类技术是一种有指导的学习(Supervised Learning)，即每个训练样本的数据对象已经有类标识，通过学习可以形成表达数据对象与类标识间对应的知识。从这个意义上说，数据挖掘的目标就是根据样本数据形成的类知识并对样本数据进行分类、进而也可以预测未来数据的归类。学习到的知识可以表述为规则，概念树，以及学习完成后的网络等形式。

神通K-Miner软件V5.0系统的分类方法包括：决策树、分类回归树、神经网络分类、支撑向量机分类、朴素贝叶斯、贝叶斯网络和逻辑回归。

### 决策树

决策树是从一组无次序、无规则的事例中推理出分类规则并以树状形式表示。它采用自顶向下的递归方式，在决策树的内部结点进行属性值的比较，并根据不同的属性值判断从该结点向下的分支，在决策树的叶结点得到结论。从根结点到叶结点的一条路径就对应着一条合取规则，整棵决策树就对应着一组析取表达式规则。

### 分类回归树

分类回归树CART（Classification and Regression Trees）是分类数据挖掘算法的一种。该模型使用了二叉树将预测空间递归划分为若干子集，Y在这些子集上的分布是连续均匀的。树中的叶节点对应着划分的不同区域，划分是由与每个内部节点相关的分支规则（Splitting Rules）确定的。通过从树根到叶节点移动，一个预测样本被赋予一个惟一的叶节点，Y在该节点上的条件分布也被确定。从根结点到叶结点的一条路径就对应着一条合取规则，整棵分类回归树就对应着一组析取表达式规则。

### 神经网络分类

人工神经网络是一组连接的输入/输出单元（即神经元）。神经元之间通过权值相联，在学习阶段，通过某种算法调整神经网络的权值，使得能够预测输入样本的正确目标值。BP算法是一种前向人工神经网络学习算法，采用BP算法训练的神经网络简称BP网络。神通K-Miner系统的神经网络分类，是预测目标值为分类类型的BP神经网络算法。

### 支撑向量机分类

支撑向量机理论，建立在统计学习理论的VC 维理论和结构风险最小原理基础之上，根据有限的样本信息在模型的复杂性(即对特定训练样本的学习精度，Accuracy)和学习能力(即无错误地识别任意样本的能力)之间寻求最佳折衷，以期获得最好的推广能力(GeneralizatinAbility)。已有许多事实证明，结构化风险最小化原则(Structural Risk Minimization, SRM )要优于传统的经验风险最小化原则（Empirical Risk Minimization, ERM）。支撑向量机的主要应用领域有模式识别、函数逼近和概率密度估计等等。神通K-Miner系统的支撑向量机分类，是预测目标值为分类类型的算法。

### 朴素贝叶斯

朴素贝叶斯学习模型（Simple Bayes 或Naïve Bayes ）将训练实例*I*分解成特征向量*X*和决策类别变量*C*。简单贝叶斯模型假定特征向量的各分量间相对于决策变量是相对独立的，也就是说各分量独立地作用于决策变量。尽管这一假定一定程度上限制了简单贝叶斯模型的适用范围，然而在实际应用中，不仅以指数级降低了贝叶斯网络构建的复杂性，而且在许多领域，在违背这种假定的条件下，简单贝叶斯也表现出相当的健壮性和高效性，它已经成功地应用到分类、聚类及模型选择等数据挖掘的任务中。

### 逻辑回归

线性回归中估计未知参数时主要采用最小二乘法，其原理是根据线性回归模型选择参数估计值，使因变量的观测值与模型估计值之间的离差平方值最小。而Logistic回归模型的参数估计采用最大似然估计方法，这一方法首先要建立一个似然函数，这一函数将观测数据的概率表述为未知模型参数的函数，模型参数的最大似然估计是选择能够使这一函数值达到最大的参数估计值。即这套参数估计能够通过模型以最大概率再现样本观测数据。

### 组合模型

组合模型，是将已经建立成功的多个分类模型按照一定的准则进行组合，形成的新模型，即分类组合模型。一般来说，分类组合模型与单个分类模型相比，预测性能更高。

## 回归

回归预测主要是对目标字段为数值型字段进行预测，是一类很重要的预测方法。本系统包括三种回归预测算法，分别是线性回归、神经网络回归和支撑向量机回归。回归预测挖掘的流程包括建模、测试、打分和实时打分。建模：对训练数据选择合适的回归预测算法进行建模；测试：选择测试数据对已建立的模型进行评估，通过一系列评估指标检验模型的性能以及模型的泛化能力，只有通过测试的模型应用于实践才有实际指导意义；打分：回归模型经过评估后可用来预测并指导实践。

### 线性回归

当我们有很多组这样的数据，这些就是训练数据，我们希望学习一个模型，当新来一个面积数据时，可以自动预测出销售价格（也就是上右图中的绿线）；这样的模型必然有很多，其中最简单最朴素的方法就是线性回归，也就是我们希望学习到一个线性模型（上右图中的红线）。不过说是线性回归，学出来的不一定是一条直线，只有在变量x是一维的时候才是直线，高维的时候是超平面。

### 神经网络回归

BP神经网络是应用较为广泛的前向传播神经网络，即所处理的信息逐层向前流动，当学习权值时，根据理想输出与实际输出的误差，由前向后逐层修改权值。它采用光滑活化函数，具有一个或多个隐层，相邻两层之间通过权值全连结。神通K-Miner系统的神经网络回归，是预测目标值为数值类型的BP神经网络算法。

### 支持向量机回归

支撑向量机理论，建立在统计学习理论的VC 维理论和结构风险最小原理基础之上，根据有限的样本信息在模型的复杂性(即对特定训练样本的学习精度，Accuracy)和学习能力(即无错误地识别任意样本的能力)之间寻求最佳折衷，以期获得最好的推广能力(GeneralizatinAbility)。已有许多事实证明，结构化风险最小化原则(Structural Risk Minimization, SRM )要优于传统的经验风险最小化原则（Empirical Risk Minimization, ERM）。支撑向量机的主要应用领域有模式识别、函数逼近和概率密度估计等等。神通K-Miner系统的支撑向量机回归，是预测目标值为数值类型的算法。

### 广义线性回归

## 聚类

聚类，是把数据按照相似性归纳成若干类别，使得同一类中的数据彼此相似，不同类中的数据相异。聚类分析可以建立宏观的概念，发现数据的分布模式，以及可能的数据属性之间的相互关系。本系统提供了三种聚类算法，即K均值聚类、分布估计聚类和层次K层次聚类算法。

### K均值

K均值是一种基于划分的聚类方法，以k为参数，把n个对象分为k个簇，使得类内具有较高的相似度，而类间的相似度最低。相似度的计算根据一个簇中对象的平均值（被看作簇的中心）来进行。K均值算法首先选择k个对象，每个对象初始地代表了一个簇中心。对剩余的每个对象，根据其与各个簇中心的距离，将它赋给最近的簇，然后重新计算每个簇的平均值。这个过程不断重复，直到准则函数收敛。

### 分布估计聚类

分布估计聚类，是基于多元高斯混合模型的统计聚类方法，该算法假定每个样本群体满足一个特定的分布，而任何一个分布均可以表示成几个高斯分布的组合。神通K-Miner系统的分布估计聚类算法可自动确定聚类簇数，也可通过用户指定最终簇数。

### 层次K均值

层次K均值是一种以K均值聚类算法为核心的划分聚类方法，通过初始建立k均值模型，再基于分群结果采用自下而上的思想，将分群个数逐渐减少，从而形成一棵层次聚类树。

## 关联分析

关联分析，用于反映一个事件和其它事件之间的依赖或关联关系。数据库中的数据关联是现实世界中事物联系的表现。数据库作为一种结构化的数据组织形式，利用其依附的数据模型可能刻画了数据间的关联。但是，数据之间的关联是复杂的，不仅是上面所说的依附在数据模型中的关联，大部分是蕴藏的。关联知识挖掘的目的就是找出数据库中隐藏的关联信息。关联可分为简单关联（超市购物篮分析）、时序关联（序列模式）、因果关联（属性关联）、数量关联等。这些关联并不总是事先知道的，而是通过数据库中数据的关联分析获得的，因而对商业决策具有新价值。

## 时间序列

时间序列，根据系统观测到的序列数据，通过曲线拟合和参数估计的方法，建立数学模型，从而进行预测，是一种被广泛使用的预测方法。时间序列要求其目标分析字段为数值型变量；另外，与其它回归算法不同，时间序列要求的数据必须按照某一个时间尺度排序，所以还必须指定一个取值唯一的字段作为时间字段，从而唯一标识单条记录。