第四部分　　项目技术规范和服务要求

项目编号：

采购单位（采购人）名称：杭州市滨江区社会发展局

一、项目概况

1.1 项目背景

随着全球信息化飞速发展,“云计算”和“大数据”时代的到来,传统的网络和计算架构已经不能很好的适应“大数据”高速响应的要求。开启大数据平台建设对于解决医疗信息化的瓶颈问题,推动医疗信息化向深度和广度迈进,进一步提升卫生部门核心战斗力有着极其重要的意义。

为了更好的管理辖区内医疗机构医疗运营情况,将医疗卫生数据进行统一管理,达到模式先进、流程优化、管理配套、支撑有力、运作高效;实现医疗信息的管理一体化,实现各下属医疗机构日常业务管理、临床医疗管理、医院资源管理、控制管理的信息化和网络化;促进医疗卫生管理和机制创新,促进管理和决策更加科学,提升卫生系统信息化素质,使社发局及下属基层卫生医疗机构在现代化管理方面处于领先地位,取得更好的社会与经济效益。

滨江区社发局在2015年开始了区域卫生信息化建设,依托云平台租赁模式开展了智慧医疗项目建设,进行软件建设和所需硬件的集

成。包括，区域卫生信息平台、区域HIS、检验、体检、PACS、PASS、健康滨江区app、智慧医疗自助服务系统、统一支付平台、综合管理平台等系统，实现区域内3家社区卫生服务中心及35家社区卫生服务站的信息集成。

1.2 项目总体目标

本方案将依托滨江区社发局基层医疗卫生机构业务数据构建底层大数据中心，为后续各级平台应用提供数据基础与保障支撑，也是滨江区基层医疗政府补偿机制改革的重要组成部分，主要包括业务系统数据中心、数据集成、数据安全三部分建设内容。

功能区块	产品名称	描述	数量
数据中心	数据中心建设	大数据中心是滨江区智慧医疗建设的基础，存储医疗卫生机构各应用系统提供的各类的数据。主要建设内容包括数据中心建模、数据中心搭建以及数据中心ETL设计。	1套
	数据中心备份	数据库实时备份管理平台，实现数据库级别的实时备份，集中WEB控制台统一管理备份节点，实现策略配置、集中监控、调度运行、一键式操作等任务。	1套
	自动化运维	数据库自动化运维平台，实现数据库智能化运维，提供数据库自动化巡检、性能分析、	1套

		运行趋势分析、健康评分和故障诊断等功能；同时实现数据中心运行状态的一体化监控，包括中间件、数据库、操作系统、虚拟机、物理服务器、存储设备、网络设备、安全设备等软硬件节点。	
数据集成	数据同步工具	实现数据库同步，和数据支撑平台结合，作为数据支撑平台的实时数据挖掘客户端，集中 WEB 管理，实现策略配置、 集中监控、调度运行、一键式操作等任务。	1 套
	数据支撑平台	一款自行研发的数据抽取转换和加载工具，可以实现按秒、分、时、天为间隔的数据抽取操作。将杂乱无章、多源异构的业务数据识别出来，并建立中心数据库的模型，将融合后的数据装入大数据中心。所有的数据从源端到目标端的流转都可以进行监控，进行数据质量保障。	1 台
数据安全	数据脱敏	数据脱敏系统结合客户各种应用场景的需求，对敏感信息通过脱敏规则进行数据变形，漂白和替换；包括：高度隐私数据进行信息置换脱敏；保持数据特征、业务规则、数据关联性，确保开发测试等场景正常使用；	1 套

1.3 项目建设原则

1、整体规划、分步实施。

按照我区智慧医疗建设的总体部署和要求，结合实际情况，进行详细调研分析，避免重复建设、资源浪费，从整体上规划安排，分步骤、分阶段稳步实施。

2、立足应用、前瞻设计。

以需求为导向，充分考虑医疗卫生机构现实业务需要，同时采用国内先进系统架构理念和技术，兼顾与公安、人口、社保等其它系统，以及与省市卫生信息平台双向交互，满足扩容和集成需求，为今后发展及系统升级留有空间。

3、资源共享、保障安全。

以现有网络、信息资源和业务系统为基础，加强网络资源与信息资源整合，实现信息资源共享。在保证信息系统先进性的同时，确保业务处理系统的安全性，数据信息资料的完整性、可靠性。

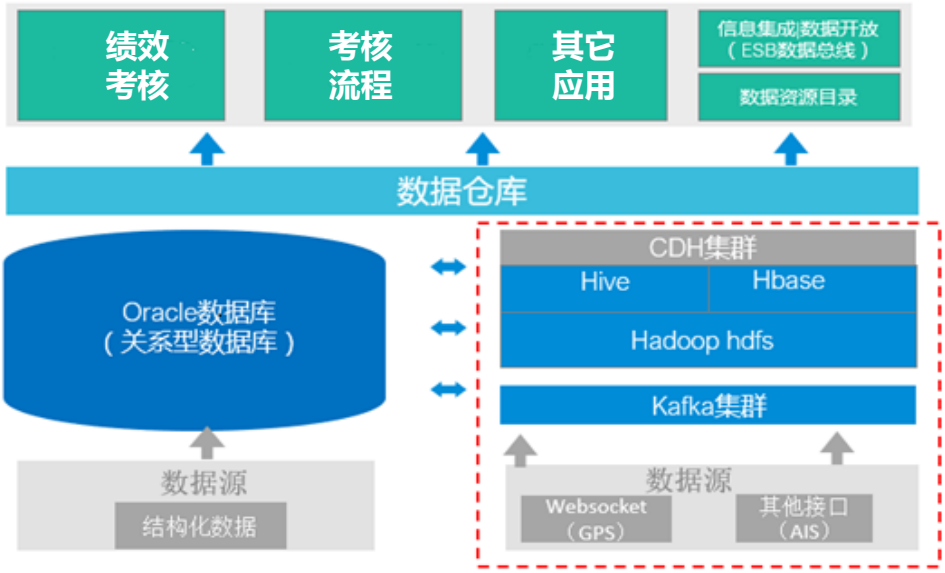
二、建设内容

序号	项目名称	数量
1	数据中心建设	1 套
2	数据库实时备份系统	1 套
3	自动化运维系统	1 套
4	数据支撑平台	1 项

5	数据同步工具	1 套
6	数据脱敏系统	1 套

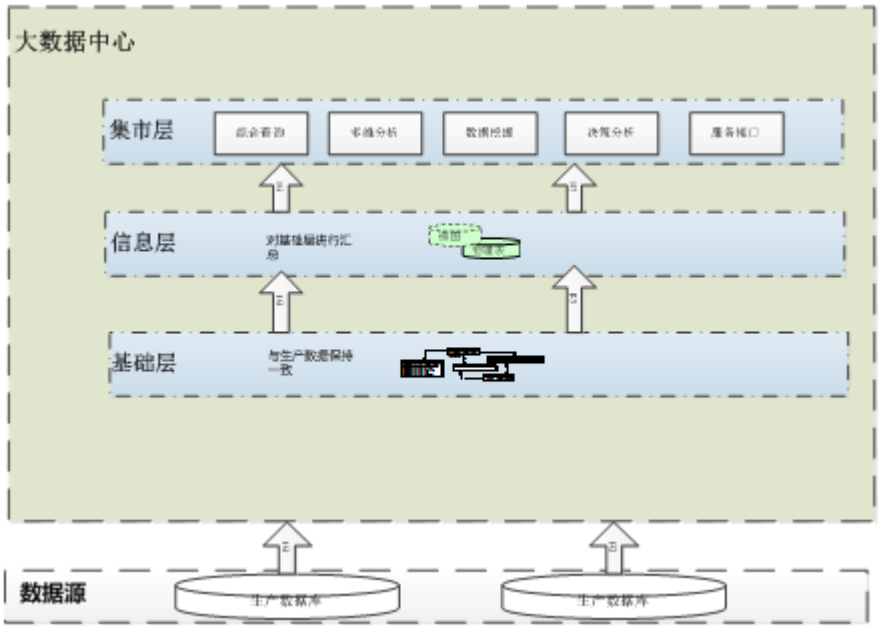
一、数据中心建设

大数据中心是医疗卫生机构财政补偿机制绩效考核的基础，存储医疗卫生机构财政补偿机制绩效考核所需要掌握和对外提供的各类的数据。数据中心建设几乎涉及所有的数据资源类型，主要包括：HIS、LIS、体检、PACS/RIS、PASS、心电、EHR、EMR、全科医生医养护签约、计划免疫以及其它基本公卫等信息系。因此，需要从满足实际应用的角度，应采用多种数据混合存储方式来进行大数据的存储设计。



整个数据中心按标准进行分层规划，标准化搭建

1）分层架构图



分层列表说明:

数据层次		定位	功能描述	数据结构	数据来源	数据粒度
应用层	决策支持	多维分析数据钻取数据不可直接修改非实时、历史性数据分析	为 KPI、报表和多维分析等决策支持类应用提供多维数据及界面展现需要的实例级数据。	星型或雪花	信息层数据	多维度汇总数据 少量实例级汇总数据
	业务运营	清单类数据提取	为针对性营销、自助取数等辅	关系型数据结	信息层数据	多维度汇总数据

		多维分析系统可能产生、修改数据 可能需要实时性数据	助业务运营类应用提供用户清单及深度分析数据。	构 业务宽表		实例级汇总数据
信息层	轻度汇总	大数据量轻度汇总，提高性能 数据不可直接更新	为了提高系统处理性能，对大数据量或逻辑处理复杂的数据进行轻度汇总。	星型或雪花模型设计	基础层数据	比实例级更详细的汇总数据
	标准数据	数据标准化处理	对于需要标准化的表进行数据标准化。	三范式	基础层数据	实例级详细数据
基础层		存储最细层历史数据	存放完整详细历史数据 ，为后续的整合数据层提供灵活性和扩展性的	三范式	数据源数据	实例级详细数据

			基础，不涉及跨主题域数据整合			

2）基础层

基础层模型不涉及跨主题域数据整合,其数据结构也基本和源系统保持一致，存储最细粒度的历史数据，所以在设计基础层模型时需要注意以下事项:

数据结构符合三范式规则，减少数据存储压力;

资料类信息表采用历史表，记录详细的历史变化轨迹，以方便交易、事实数据关联资料类信息表可以根据时间戳准确还原历史数据场景。

实体主题域归属、表命名、字段命名参考附录中的数据仓库相关规范;

3）信息层

信息层标准化模型是对基础层最细粒度的实例数据进行数据质量标准化处理，所以在设计信息层模型时需要注意以下事项:

模型和基础层基本一致，但是需要进行表名标准化、字段名标准化、类型标准化 等一系列模型标准化处理。

数据标准化，需要定义数据标准，类型数据统一，数据列为引用数据按照引用标准进行标准化处理。

基础层数据对象一致，但是是分散的，如多个不同医院的同一个

模型数据，如多家医院门诊就诊信息表，按照标准进行数据汇聚，放到同一个标准表中，便于后续数据使用。

信息层汇总模型是对基础层最细粒度的实例数据进行轻度汇总，所以在设计信息层模型时需要注意以下事项:

数据结构尽量采用星型模型，适当冗余部分常用维度描述信息，提高模型的可理解性，同时满足用户对数据查询性能的要求;

部分常用的复杂业务逻辑的口径直接封装为标识类字段;

资料类信息表采用历史表和当前表相结合的方式，既能满足对历史信息的存储，又能满足当前最新信息的跟踪;

交易、事实类数表采用日表和月报相结合的方式，记录不同时间粒度的汇总数据，降低日表的访问压力;

信息层模型往往涉及跨多个源系统数据的整合,不同系统的编码、数据类型等需要有一个统一处理的过程。

4）集市层

应用层模型，是为分析主体进行构建。

应用展现层模型一般都是汇总数据，不到实例级，通常清单级的数据我们一般都应该在信息层封装。应用层模型往往和具体报表有着密切的联系，或者说通过一定的维护汇总就是一张报表，所以在设计应用层模型时需要注意以下事项:

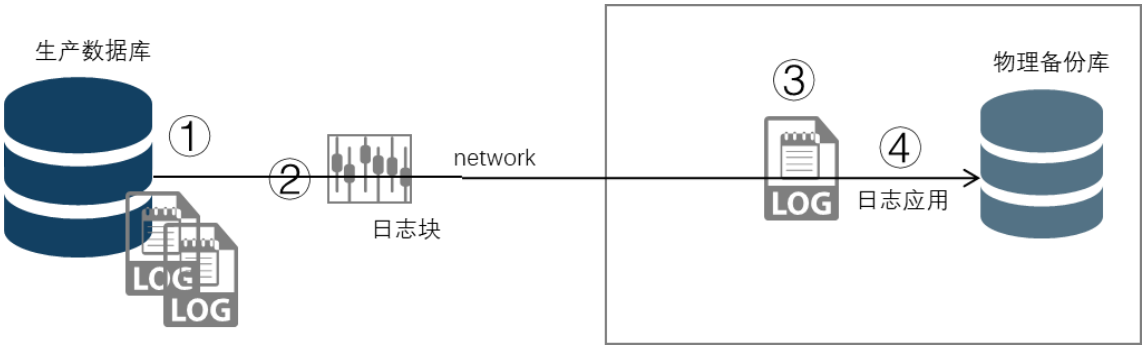
维度和指标通常不是绝对，在特定场景下是可以相互转化的，例如一个需求需要统计网上用户,我们可以直接将其作为一个指标网上用户数，也可以作为一个维度网上用户标识。但如果在加一个条件统

计多个客户下的网上用户数，相对而言我们在设计的时候将其设计成指标比较合适

应用实体在设计时最好不要随意加维度，维度的增加会使实体的数据稀疏性、数据量都发生一定的变化。

二、数据中心备份

为了保障数据中心建设后数据完整性和备份，针对数据中心建设一套实时备份系统，在华通云上配置另外一套系统和数据库环境，将数据中心数据备份一份到该环境中，部署架构如下：



利用数据库实时备份软件在主备两端实现数据库同步，在生产库服务器上部署 Agent，用来挖掘日志文件，然后把日志文件以片段的形式传输到备库，进行日志合成和分析，从而获取数据变化信息，实现数据中心实时备份，由于数据库实时备份软件是以片段的形式传输数据，并且只传输变化的数据，可以在有限的带宽上实现远距离备份，确保两端数据的高度一致性和实时性。

容灾目标

1) 实现数据库级别的监控备份状态，提高系统运行的稳定性和可靠性。

2) 配置至少 1 个数据库实时备份许可（Oracle、Sql server 等），含一年 7*24 小时标准服务。

3）支持数据库版本：ORACLE 9i、ORACLE 10g、ORACLE 11g、ORACLE 12C、Sqlserver2008、Sqlserver2012、Sqlserver 2014、Sqlserver 2016 等。

网络带宽：数据库实时备份系统可以运行在高带宽或低带宽网络环境下，对于网络带宽不需要有太高的要求，可以运行在百兆网络、千兆网络、光纤等。

网络质量：数据库实时备份系统可以运行在稳定网络环境或不稳定网络环境下，对于网络质量不需要有太高的要求，系统运行过程中，网络出现短暂的网络终止等异常问题，灾备系统仍能保持正常运行、不会被终止。

网络容错：可以利用多块网卡或多条链路实现链路容错，当某一链路出现故障、暂停或时断时续的现象时，数据库实时备份系统可以自动识别，不能影响系统的不间断运行。

逻辑错误传播：要求避免逻辑错误传播，从而对于逻辑失败形成根本性的保护。

★数据完整性：实现所有数据库内的数据操作的复制，包括 INSERT\UPDATE\DELETE、DDL 操作、Create table 。。as 语句，ROWID 相关语句等。(投标时要求提供功能截图)

实现所有数据库内的所有对象复制，包括普通表格、压缩表格、临时表格、垃圾箱内的闪回表格等。(投标时要求提供功能截图)

要求两个库之间的所有对象完全相同，包括 ROWID，基表，视图，同义词，自定义 TYPE，IOT 表格，SYS 用户内的所有对象等。

对系统性能影响：对系统性能影响小于百分之五。

统一管理：提供直观的简体中文图形化 WEB 操作控制台，可以在一个 WEB 图形界面里管理所有业务系统的备份系统管理，要求所有备份功能统一在 WEB 界面上操作。

在线实施要求：在线实施，实施期间生产系统不得停机，业务不能终止。

在线软件升级：对备份平台进行软件升级安装时，无需停止备份平台，不中断备份业务系统的运行。

★数据库同步方式：要求采用物理同步方式，无需考虑复杂的内部数据关系，支持数据库中所有对象的同步，支持所有 DDL、DML 等语句的复制。(投标时详细描述数据同步实现机制)

虚拟化环境支持：通过容灾系统可以自动实现虚拟化环境上的各类数据库的同步。

★对业务系统影响：要求备份系统对生产业务系统无影响，无需改造生产系统主机上的文件系统，不能更改系统卷配置，生产数据库内不允许嵌入任何程序，不修改生产数据库配置参数，不增加数据库内的表格或其他对象。

备份能力：要求备份站点可实现常规物理备份，以代替生产系统备份。

备份兼容性：备份系统的一体化备份要求被 RMAN 兼容，可以

使用 RMAN 完成数据库恢复。

备份系统在线校验：实现备份系统在线校验，支持备份系统运行时随时打开容灾库以检查备份可用性。

API 二次开发接口：提供 API 接口，具备二次开发能力，能够定制部分备份平台功能。

短信通知：提供手机短信预警接口。

★拓展服务：要求备份站点数据库可以作为逻辑备份、数据仓库等源数据，被应用程序或 ETL 工具所使用。

数据更新能力：要求备份站点实现边同步边查询的能力，活动数据库提供活动数据而不是静态数据，随时跟踪生产站点数据库，差异性间隔时间：0 至 5 分钟。

数据安全保护：备份系统支持备份数据安全保护，对备份数据中的敏感数据进行安全保护，防止未经授权的工具和应用或未经授权的用户随意访问。

三、数据库自动化运维

运维系统是主动运维服务的线下服务平台，可以单独使用，也可以作为运维云的线下服务终端。运维服务系统实现数据库智能化运维，提供数据库自动化巡检、性能分析、运行趋势分析、健康评分和故障诊断等功能；同时实现数据中心运行状态的一体化监控，包括中间件、数据库、操作系统、虚拟机、物理服务器、存储设备、网络设备、安全设备等软硬件节点。运维人员利用运维服务系统可有效提升对数据

中心的自动化运维水平，实现对数据中心的主动运维服务。主要功能包括：

（一）数据库智能化运维。

（二）数据中心一体化监控。

（三）实现告警订阅、工单管理、知识库和远程 DBA。

（四）大屏展示，可自主定制运维视图，实现运行状态可视化。

一、总体架构

系统架构：采用 B/S 架构，支持中文管理界面。

部署方式：★支持集中式和分布式部署方式，实现跨机房的数据中心运维监控。

★监控对象采用无代理方式部署，支持 SNMP、SSH、JMX 等采集模式，采集过程要求不影响监控对象的性能。

可以作为数据中心服务平台单独使用，也可以作为运维云的线下服务终端使用。

二、数据中心一体化监控

★中间件监控：支持 WebLogic、Tomcat 等主流业务中间件监控。采用 JMX 直接监控 Java 应用程序服务器的功能，无需第三方模块或集成层。使用高效的 Java 网关监视 Tomcat 等应用服务器，包括：JVM 可用内存、JVM 最大内存、JVM 总的内存、线程等。(投标时要求提供功能截图)

操作系统监控：支持 Linux、Windows、AIX、HP-UX、Solaris 等操作系统监控。监控操作系统的运行状态，包括：主机状态、CPU

利用率、内存利用率、磁盘 I/O 读写速率、磁盘 I/O 读写速率，网络 I/O 等。

★虚拟机监控：支持 VMware 虚拟机状态监控，包括：VMware 物理主机硬件状态、虚拟机在线状态、CPU 利用率、内存大小及利用率、磁盘空间大小及利用率等。(投标时要求提供功能截图)

服务器监控：支持 IBM、HP、DELL、联想、曙光、浪潮等主流物理服务器监控，包括对服务器的硬件状态进行自动告警，包括：CPU、硬盘、电源、风扇、RAID 卡等。

存储监控：支持 EMC、NetApp、IBM、华为等主流存储设备监控，包括：存储系统的磁盘、存储控制器、电源、风扇等运行状态。

网络监控：支持华为、H3C、思科、锐捷等主流网络设备监控，包括监控网络设备的在线状态。对网络线路运行状态监控，包括线路连通性、线路响应时间、线路流量、线路带宽利用率、线路错包率、线路丢包率等信息。对网络设备接口状态进行监管，包括接口状态、接口流量性能等信息。

三、数据库监控

数据库类型支持：支持 Oracle、SQL Server、MySQL、DB2 等数据库。

状态监控：支持数据库运行健康状态监控，包括：

可用性监控：监听、实例、表空间的可用性；

错误监控：数据库运行过程中的错误数量；

性能监控：数据块逻辑读指标直观反映数据库性能；

变化监控：对象（表、索引、视图等)、权限（用户、表)、空间（对象、表空间、归档）的变化量；

可靠性监控：备份及容灾系统的运行状态。

★SQL 执行监控:支持 SQL 执行生命周期完整监控,包括:登录、解析、执行、提交。通过 SQL 执行次数、SQL 执行时间，主观展示 SQL 执行性能瓶颈。(投标时要求提供功能截图)

关联资源监控：支持数据库关联资源监控，包括：

1）数据库资源监控：processes、session、DB files、jobs。

2）三大资源锁监控：Mutex、 Latch、 Lock。

3）主机资源监控，包括：CPU、内存、存储、网络。

★数据库巡检：需要对巡检对象的业务系统、数据库类型、实例名、操作系统、IP 地址、巡检完成时间等信息进行全面巡检。

支持巡检对象的异常信息数量统计，运维人员可及时掌控各系统数据库的健康状况。

整合资深数据库专家的多年运维经验，形成专家智能系统，根据数据库的健康状态，给出专业的分析，对数据库可能存在的故障和问题进行快速定位，对每个异常指标提供专业解读和排错建议。

1、支持一键操作实现全面巡检

2、全面巡检的报告内容至少包括：数据库可用性、空间管理、安全性、可靠性、性能、错误、主机资源、数据库资源、数据库软件、数据库参数、系统参数等信息。

3、支持在线和以 PDF 文档导出两种方式查看巡检报告。(投标时

要求提供功能截图)

四、告警触发模式

人工阀值模式:支持人工阀值模式,可自由设置告警触发阀值库。

五、大屏展示

★预封装模版：预先封装不同对象的大屏展示模版，包括：中间件、数据库、操作系统、虚拟机、物理服务器等。模版的内容支持:一张大屏可以直观显示监控对象的健康指数，并根据阀值自动告警。从流程及时间模型的角度联动展示各项指标，清晰定位问题环节。常见运维故障（80%以上）及隐患能从大屏中直观展示及告警。(投标时要求提供功能截图)

自主定制：支持按照业务角度进行关联视图定制。支持按照资源种类进行分类和关联视图定制。支持机柜图、系统拓扑图、网络拓扑图等视图定制。

六、运维云

★告警订阅:可定义提交的告警与技术服务人员接收的对应关系,包括：告警站点、告警级别、通知方式等。(投标时要求提供功能截图)

运维数据查询：采用大数据结构的集中管理模式，可高效查询、分析。

1）支持告警日志查询、分析。

2）支持状态性能数据查询、分析。

3）支持历史巡检报告、性能分析报告查询、分析保存 3 年以上运

维历史数据。

工单服务：提供工单流转、统计、报表等功能，与告警订阅和服务授权紧耦合；针对“待处理告警”，支持一键发起服务工单。

告警通知：支持微信、App、邮件等通知方式

★工具体验：提供数据库巡检和 AWR 报告解读工具，支持生成巡检报告和 AWR 解读报告

远程数据库专家服务：授权方可选择指定类型告警同步推送给远程数据库专家，数据库专家结合告警和运维数据快速、高效地交付服务。结合运维平台提供的巡检报告和性能报告，可主动地提出优化数据库性能建议

七、产品规格与授权

授权数量：配置至少 6 个数据库实例监控许可，100 个网络服务器存储等其他设备监控许可。

四、数据集成

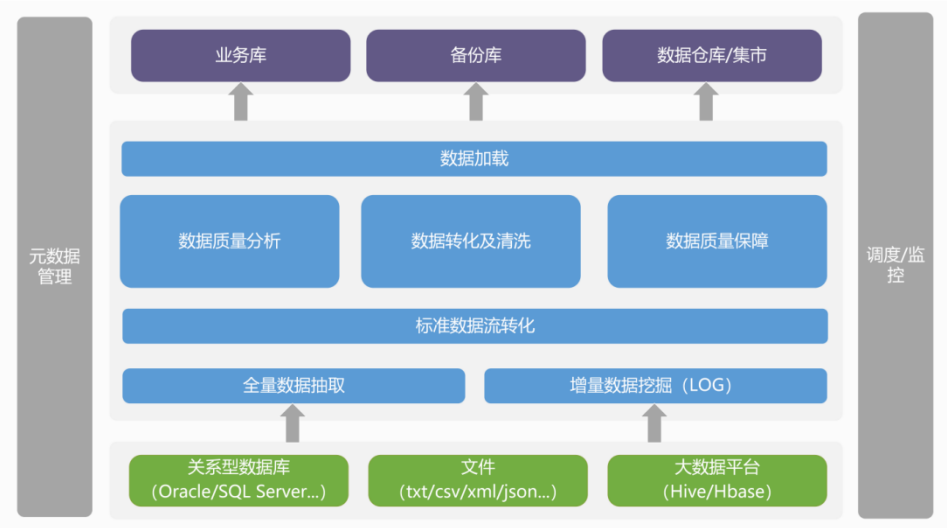
数据集成是将多源异构的滨江区社发局业务数据抽取、转换、加载至数据中心的过程，在数据挖掘过程中，数据集成是第一步骤，即对数据进行预处理的过程。各种不同的挖掘系统都是针对特定的应用领域进行数据集成的。在开始进行数据转换、采集的工作之前，首先要对数据的现有情况进行分析，数据转换、采集清洗的工作是对现有的实际数据进行的操作。

由于不同系统提供的数据内容、数据格式和数据质量千差万别，甚至会遇到数据格式不能转换或数据转换格式后丢失信息等棘手问

题，严重阻碍了数据在各部门和各应用系统中的流动与共享。因此，很多单位需要一个可以上述问题的统一的数据集成工具。

五、数据支撑平台

数据支撑平台主要包括数据抽取工具、数据清洗工具以及数据质量治理三部分，从多个方面解决项目中的数据集成问题。其中数据抽取工具可以定时将数据源中的增量数据取出，并按照指定格式（如XML）输出数据；数据清洗工具主要把垃圾和异常数据清洗掉、数据转化；数据质量治理针对主数据，提供数据质量保障。



总体架构图

数据支撑平台系统提供的最主要的功能单元包括：

➢ 数据抽取单元：从数据库或外部文件中读取数据，支持增量抽取。

➢ 数据处理单元：实现数据的拼接、数据分拆、数据过滤、数据条件、数据复制、数据排序、数据计算、数据转换等各种功能。

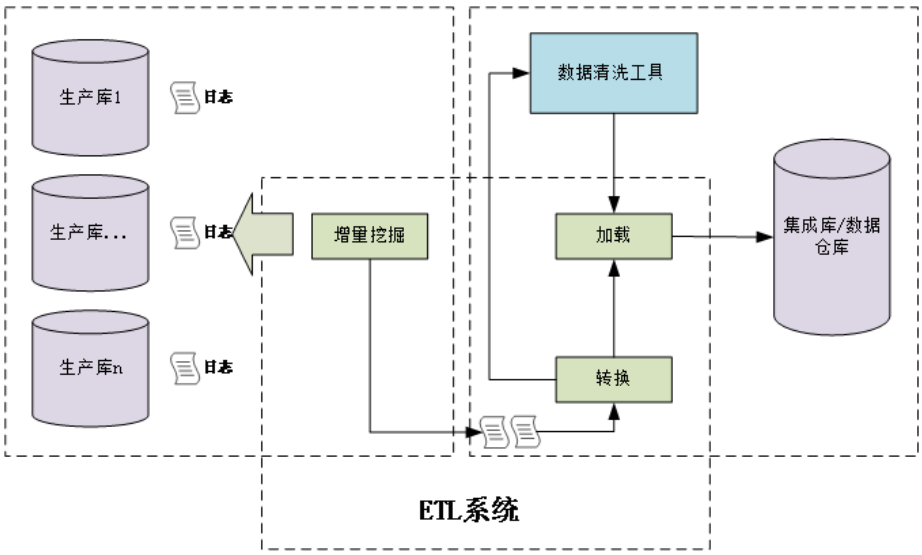
➢ 数据加载部件：支持以 SQL 语句的方式更新目标数据库中的数

据、以外部文件批量装载的方式装载数据到目标数据库中或者将数据导出到外部文本文件。

数据支撑平台系统不仅包含了同类产品中常用的功能组件和转换函数，更增加了数据质量检测、清洗的函数以及主键转换函数等丰富功能，能够满足用户数据处理的多种需要，并能提高数据处理的质量。

1）增量采集

数据支撑平台中数据抽取转换工具采用增量采集技术，增量采集系统对生产库的日志文件进行增量监控，有变化就把增量数据采集并送到 ETL 工具端，经过 ETL 一系列转换和清洗后，最终数据以目标端的组织形式入库。日志监控的增量采集方式对生产产生的干扰少，同步快，也有更高的稳定性。



2）多源支持

数据支撑平台中数据抽取转换工具支持丰富的输入源，使得用户

可以方便高效的从各种来源抽取想要的数据。具体的输入格式如下:

3）数据转换

数据转换的任务主要是针对不一致的数据按规则进行转换，进行数据粒度的转换，以及一些商务规则的计算。

1、不一致数据转换：这个过程是一个整合的过程，是指将不同业务系统的相同类型的数据统一，比如同一个供应商在结算系统的编码是 XX0001，而在 CRM 中编码是 YY0001，这样在抽取过来之后统一转换成一个编码。

2、数据粒度的转换：业务系统一般存储非常明细的数据，而数据仓库中数据是用来分析的，不需要非常明细的数据。一般情况下，会将业务系统数据按照数据仓库粒度进行聚合，从而为后续的分析工作打下基础。

3、商务规则的计算：不同的企业有不同的业务规则、不同的数据指标，这些指标有的时候不是简单的加加减减就能完成，这个时候需要在 ETL 中将这些数据指标计算好了之后存储在数据仓库中，以供分析使用。

数据支撑平台中数据抽取转换工具支持丰富的转换功能，能适应各种数据转换场景，转换功能包括普通数据转换、应用类转换和流程类转换。

支持扩展类转换:JAVA 代码组件、JS 代码组件、Rule Accumulator、公式、Janino、SQL 脚本、正则表达式等。

支持数据转换和数据清洗的一体化配置。

4）数据治理

数据支撑平台是专业的数据质量分析,比较,验证和监督的软件,它会对各系统的企业信息进行数据分布验证（如代码、名称的分布情况），基于验证结果，提取关键信息进行比较，转换并提炼唯一的标准企业名称,并保证之后新抓取的数据都会根据已经设定的数据质量治理流程进行标准化处理。主要实现了以下功能:

1、数据源选择和数据转换条件模块,使用户可以方便的从界面选择;

2、数据表浏览模块和数据表信息统计模块,使用户可以在数据清洗过程中可以掌握数据表的信息，从而指导下一步的数据清洗;

3、包括蕴含值分离，分离字段合并，字符型字段规范化，缺失值处理，重复元组检测，异常值检测等等各种数据清洗算法，可以很好地完成数据清洗任务;

4、自动清洗向导,极大地方便了对数据清洗缺乏足够了解的用户;

5、提供了可视化界面将清洗结果返回给用户,用户可以对清洗结果进行各种处理，当然也可以选择系统自动处理。

- 选择数据源

通过这个模块选择要进行清洗的数据集。用户通过点击下拉框可以获得所有在机器上注册的数据源名称,选定一个数据源后又可以获得该数据源中包含的所有表。当用户选定要操作的数据表后,这个表中所有的列都会显示在界面上。用户可以选择这个表所有的列,也可以选择其中感兴趣的几列进行清洗。

- 提取元数据

元数据为访问数据仓库提供了一个信息目录(information directory)，这个目录全面描述了数据仓库中都有什么数据、这些数据怎么得到的、和怎么访问这些数据以及这些数据的基本信息。数据仓库服务器利用他来存贮和更新数据，用户通过他来了解和访问数据。而对数据进行清洗时，应同时对元数据库进行修改，使当前元数据库的存放的是最新的信息，以便提供给数据集市和数据立方操作必需的信息。

- 选择目的表

用户在这个模块中可以选择目的数据源，目的数据表名。这是由于在数据清洗的过程中，需要对数据集中的数据进行一系列更新、插入、删除等操作。这些操作都是无法恢复的。为了保证在数据清洗结果不理想的情况下可以恢复成原来的数据，我们不在源数据集上直接进行数据清洗操作，而是生成一个与源数据表相同的目的数据表。这样我们就可以在这个复制的表上进行数据清洗，如果数据清洗的结果满意的话，我们再对源数据表进行相同的数据清洗操作或者直接用目的表替换源数据表，这样就保证了源数据的安全。

- 数据转换条件选择

在这个模块用户可以通过输入特定转换条件来选择是对源数据表中所有记录进行清洗还是只对一部分数据进行清洗。这是因为有的时候用户可能不想对一个表中所有的数据进行清洗，而只是想对某些满足特定条件的记录进行清洗。

- 数据集信息统计

这个模块的功能是计算给定数据集的各项统计数据。比如说，每个字段的均值，方差，空值个数，不同值个数，最大值，最小值等。用户可以在其他的各个模块中调用这个模块,这样用户就可以随时获得数据集的统计信息，从而更好的进行数据清洗操作。

- 浏览数据表

用户可以通过这个模块浏览源数据表,目的表,以及清洗结果表。并可以对数据表进行更新，删除等操作。

- 手动选择数据清洗算法

该模块是通过人机交互方式，采用不同的算法处理数据集中的不规范数据，异常数据以及重复元组，得到推荐的异常数据值和重复元组,并同检测前的数据相比较,由用户决定如何处理,若用户选择“系统自动”处理的话，则根据算法自动处理异常值和重复元组。具体包括以下几个子模块:

1) 对字符型数据进行规范化处理:这个子模块主要处理字符类型的字段中不规范的数据格式，比如说不对称的引号，乱码，多余的空格等。

2) 去除空值:该子模块找出数据集中的空值，根据用户选择的方式填充空缺值。

3) 分离字段:该子模块主要处理内含值问题。内含值问题是指本应属于多个字段的值却出现在一个自由格式的字段里。在这个模块中用户可以通过指定分隔符或者直接划定的方式将其分离开来。

4) 合并字段：在数据库同样存在着这样一种问题，即本应属于同一个字段的数据却错误的被划分到两个不同的字段中。这个子模块可以通过合并两个列来解决这种问题。

5) 检测数据是否越界：这个子模块主要是用来处理数值超出字段范围的情况。比如说，无效的工资，时间或者电话号码数据等。举例说明，Salary=界．10l。这就需要用户根据对字段的了解指定阈值检测数据是否越界，把越界的数据作为异常返回给用户，有用户进行相应的处理。

6) 根据函数关系检测异常值：有的列与列间存在着一定的函数关系，比如说，某一列可能表示前面几个列的和或者均值等。还存在这样一种情况，就是一系列属性间不满足一些特定的规则，如年龄和出生日期不符等。我们可以根据用户输入这些规则或者函数关系，然后利用这些函数关系和规则进行检测，对于不满足的记录作为异常值返回给用户。

7) 根据统计数据检测异常值：该子模块计算出每个字段的各种统计值，然后根据这些统计值检测数据，选择偏离最大的前 n 个记录作为异常返回给用户。

8) 根据关联规则检测异常值：该子模块根据用户给定的阈值找出数据中满足的关联规则，再根据这些关联规则找出异常值。

9) 重复元组检测

- 自动清洗向导

由于对数据集进行数据清洗需要一些专业知识，一些用户可能对

数据清洗的流程不太清楚。为了解决这种情况本系统提供了自动清洗向导模块以向导的方式一步一步的指导用户完成数据清洗任务。具体的步骤是:

1) 提示用户选择字符类型的字段进行规范化处理;

2) 找到各个字段中的空值并将其去除;

3) 将某个字段分离成两个字段;

4) 将某两个字段合并成一个字段;

5) 通过用户限定最大最小值找出异常值;

6) 根据各个列间的函数关系找出异常值:

7) 根据统计数据检测异常值;

8) 根据关联规则检查异常值;

9) 对数据集进行重复元组检测。

5）作业调度

数据支撑平台中数据抽取转换工具的作业调度非常灵活，主要支持:

1）按时间调度:具体日期(某一天)可以是每周的某一天，也可以是每月或者年，周期(间隔时间)可以隔周、天、月、时、分、秒。

2）按外部条件调度:检查文件状态、检查服务状态、执行系统调用根据返回值调用、检查数据库连接等。

3）可以执行外部作业:系统调用、网络调用、文件传输、SSH、远程登录等。

6）监控告警

数据支撑平台提供完善的平台性能与作业执行情况的统计与监控。平台性能包括 CPU、内存、磁盘的实时资源使用情况以及历史的趋势统计，作业统计与监控包括目前正在执行的作业情况（作业名称、资源文件、启动时间、创建人、错误数、处理速度）以及历史作业的统计与分析。同时支持远程管理和监控 ETL 过程，可以看到实时的系统状态以及转换执行过程。

7）数据对账

平台处理通过监控告警，实时进行数据异常告警外，还提供数据对账功能，来全面数据质量保障。

数据对账，提供数据源时间段（日、月度等）记录数与数据中心数据量进行数据对账功能。对账的差异数据提供数据追溯，可方便发现遗漏数据的清单编码，并提供对账差异的自动补录功能。

六、数据同步

数据同步过程是利用数据库灾备复制软件将滨江区社发局业务数据实时同步至数据中心,首先利用捕捉进程(Capture Process)在源系统端读取 Online Redo Log 或 Archive Log，然后日志进行解析，只提取其中数据的变化如增、删、改操作，并将相关信息转换为软件自定义的中间格式存放在队列文件中，这样对生产系统不会产生任何影响。再利用传送进程将队列文件通过 TCP/IP 传送到目标系统。捕捉进程在

每次读完log中的数据变化并在数据传送到目标系统后,会写检查点,记录当前完成捕捉的log位置，检查点的存在可以使捕捉进程在中止并恢复后可从检查点位置继续复制;目标系统接受数据变化并缓存到软件队列当中，队列为一系列临时存储数据变化的文件，等待投递进程读取数据;软件投递进程从队列中读取数据变化并创建对应的SQL语句，通过数据库的本地接口执行，提交到数据库成功后更新自己的检查点，记录已经完成复制的位置，数据的复制过程最终完成。数据库灾备复制软件采用分布式设计，各进程模块相互独立，通过数据流进行交互，耦合度低。在系统部署上，各进程模块既支持独立部署模式又支持集中部署模式,具体部署架构根据运行环境负载与复制业务集成程度来决定，分为三种常规部署架构。常规架构一的部署方式是通过新增预处理服务器和装载服务器的方式,分别提高数据预处理和装载的效率，同时可降低目标数据库服务器压力;常规架构二的部署方式是通过新增预处理/装载服务器的方式，在提高中间数据处理效率的同时，可以降低目标数据库服务器的压力;常规架构三的部署方式支持在不新增服务器的情况下,系统相关进程分别部署在源数据库服务器和目标数据库服务器上。系统管理 web 应用部分单独部署，网络可达即可，华通云环境下需要开放数据库日志查询权限（首次全量同步开发全量数据读取权限，同步完可关闭)。

要求数据同步工具为集成平台的一部分,在特殊情况下数据复制软件能够独立使用，提供产品著作权。

利用数据库底层技术完成SQL语句之间的抽取解析传输和同步。

具有子集抽取、数据预处理、数据订阅、数据装载、复制配置、WEB展示、WEB管理等功能。

★配置至少 10 个生产数据库数据抽取许可，含一年 7*24 小时标准服务。

软件架构：要求数据同步工具采用先进的设计架构，在数据库日志抽取后，利用内存数据库的分布式架构对抽取的日志进行解析、订阅、装载处理；在内存数据库的预处理端对抽取的数据库日志按照事物进行解析合并。通过远程 jdbc 接口将数据应用到目标段。

软件部署要求：抽取端，预处理端，应用端要求采用模块化部署，可部署在不同机器。抽取端支持和生产主机进行分离，且不需要在生产数据库端开启其他挖掘选项，减少对于生产数据库的性能压力

支持数据库类型：要求支持多种数据库，包括 Oracle， MySQL、PostgreSQL 等。

同步复制功能要求：

1) 支持一对一，一对多，双向、多对一、级联复制、环形部署；

2) 支持数据抽取、转换、数据拆分、数据分发；

3) 支持对列的数据进行转换，包括列映射、增/删除列、列转换；

4) 支持 DML/DDL 操作复制、支持 SEQUENCE、函数、存储过程、视图、同义词、索引、应用包、用户等数据库对象进行复制；

5) 支持不同数据库之间的 DML/DDL 操作复制；

6) 支持按照 schema 方式设置复制关系，无需单表设置复制关系，支持不同源和目标端在不同的 schema 名情况下的复制；

7) 支持 ETL 功能，数据转换、数据拆分及分发；

8) 支持中文汉字内码，符合双字节编码；支持自定义数据格式的装载；

DDL 操作支持：能够实现同种数据库和不同数据库之间的 ddl 复制，且要求直接从日志中获取 ddl，不采用生产端嵌入触发器的模式实现 ddl。

★数据比对：支持表级别，用户级别的主备数据结构校验，数据内容校验，同时对于表结构发生映射后的数据也支持自动化比对。(投标时要求提供功能截图)

★数据加密与压缩：支持采用国密算法，dec，aes 等对称加密算法对挖掘出来的数据进行加密存储后再进行传输。加密后平均压缩比达到 8:1。

★数据库逻辑错误恢复要求：对于日常维护过程中的误操作，数据库同步软件支持备库对历史数据进行逻辑错误恢复，恢复级别能够做到全库级别，用户级别，表级别，对象级别。(投标时要求提供功能截图)

★自动化部署：通过 web 端配置复制关系后可将策略自动分发到源端和目标段，实现数据初始化和相关进程配置，完成整个复制链路搭建。

日志监控：通过 elk 技术收集所有源端和目标段的日志，对日志进行统一的分析和展现。

配置文件版本管理：支持复制链路配置文件的版本化管理，通过

web 可获取不同版本的配置情况。

★云平台部署：支持云平台上全自动化部署，包括主机创建和配置，数据库创建，同步链路部署;支持云平台包括阿里云，微软 azure，aws，openstack 等主流云平台。

★服务接口：所有服务和进程支持 resful 接口，可对接第三方监控管理中心

特殊订阅：支持挖掘数据对接 kafka，redis，hbase，hive，hdfs，strom 等大数据平台。

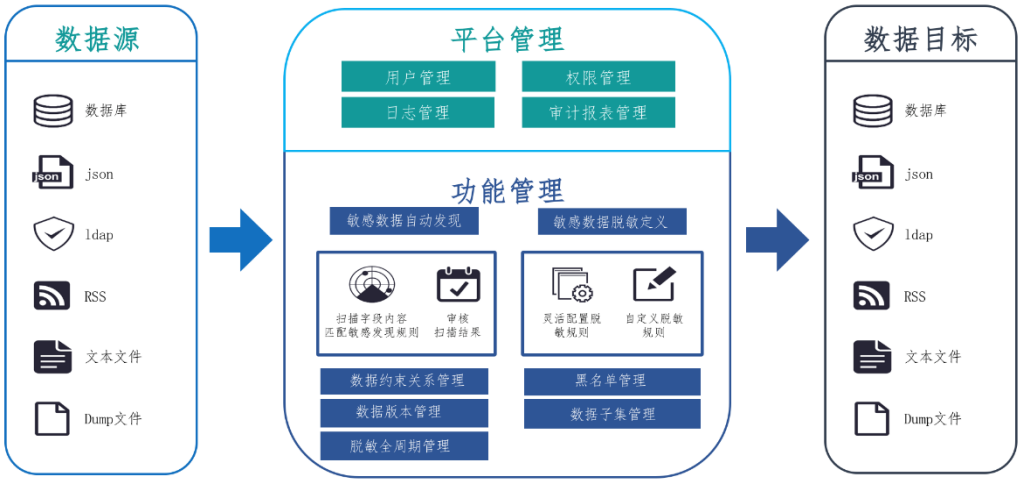
在线实施要求：在线实施，实施期间生产系统不得停机，业务不能终止。

短信通知告警：提供手机短信预警接口，邮件。

七、数据脱敏

随着区社发局各类信息系统采集、处理、积累的数据越来越多，数据量增速越来越快，数据已经渗透到滨江区社发局业务职能领域，逐渐成为重要的生产因素；而人们对于大数据的运用预示着新一波生产率增长和消费者盈余浪潮的到来。随着大数据研究的不断深入，敏感数据的概念也在不断扩张。以往认为无关紧要的数据，在大数据时代下会变的极富价值，亦将从非敏感数据变成敏感数据。

数据脱敏系统可以满足区社发局为开发环境、测试环境、培训环境等提供脱敏后的生产数据，也可于为数据交易、数据交换、数据分析等第三方数据应用场景提供适用的敏感信息泄露防护作用。



生产的敏感数据主要位于数据库中，也可能位于 csv，txt，excel 等文本文件中。脱敏系统平台主要作用是针对位于数据库及文件中的敏感数据进行有效的转换和变形，形成一份完整可用的仿真数据。同时将仿真数据自动加载到测试环境中的数据库或者文件，提供平时测试使用。

1）敏感数据发现

在敏感数据无处不在、业务越来越复杂的生产业务系统中，业务系统后台数据库表的规模越来越庞大、结构越来越复杂，数据脱敏系统利用各类敏感信息规则，通过自动扫描发现的方式高效、方便、全同的获取敏感信息，支持灵活的配置方式（包括字段信息匹配、数据信息匹配）来自动探测数据库敏感信息字段。

★脱敏系统支持灵活的敏感信息自动发现：

1、支持特定隐私敏感数据类型的自动发现功能。数据脱敏系统能够根据数据本身的特征，包括类型、长度、数据本身的编码特征、校验算法特征、语义特征等等进行数据分析、分类判断，能够分辨包含但不限于以下种类的隐私数据类型。包括：姓名，身份证，银行卡，

cvv，医生资格证书，医师执业证书，护照，军官证，护照，组织机构代码，组织机构名称，营业执照，社会统一信用代码，税务登记，开户许可证，证券名，证券代码，基金代码，基金名，电话，邮箱，ip 地址，车牌号，邮箱。(投标时要求提供功能截图)

2、支持客户根据自身特殊数据进行脱敏发现规则定义，并且可设置类型名称。(投标时要求提供功能截图)

3、根据源端数据库负载和数据质量情况，可自定义脱敏发现的抽样比例和匹配率。(投标时要求提供功能截图)

4、以报表的方式对发现结果进行展现，并可对结果进行审核。(投标时要求提供功能截图)报表中包含敏感字段（样本数据的匹配度大于等于阈值）列表。报表中显示敏感表名称、敏感字段名称、字段注释、字段类型及长度、匹配敏感数据类型；可对敏感发现的结果进行人为修正并审核。

5、在选择表格生成脱敏源的过程中，要求能够对源端表格通过关键字进行筛选，减少人工肉眼筛选的工作量。(投标时要求提供功能截图)

2）敏感数据对象预置

为满足绝大多数敏感数据识别的需求，脱敏系统设计一个敏感数据类型库，能够匹配通用的绝大部分敏感数据。

3）敏感对象脱敏规则

★脱敏系统支持灵活的脱敏规则管理：

1、脱敏系统需内置丰富的脱敏规则，包括确定随机、遮盖、仿

真脱敏（包含姓名、地址、电话、证件号码、邮箱、邮编、银行账号、组织机构代码、营业执照等）等规则；脱敏规则在界面方便管理，自定义增加（如：支持字符串和数字的映射、截取、截断、位移等）；支持依赖脱敏，即：混合字段依赖类型字段的脱敏。

2、脱敏源经过脱敏处理后，保持数据内部特征和外部关联特征，如脱敏后数据的主外键关系保留并保持一致，无主外键情况下的业务语义关联仍能保证一致。

3、对于字符型敏感数据对象，支持随机、固定映射、遮盖、截取、截断方式脱敏；对于数值型敏感数据对象，支持随机、移位、截断、取整方式脱敏。(投标时要求提供功能截图)

4、支持同一脱敏源或不同脱敏源中的相同数据，经过多次脱敏后结果一致。

5、同一列数据中包含多种数据类型(如证件号码列中包含身份证、军官证、护照)，脱敏后仍然能够生成和对应源数据一致的数据特征。

6、对敏感数据进行多个分段处理，对指定的分段进行脱敏，其余保留原指。 如身份证可以分为地址码（省）、地址码（市）、地址码（区）、出生日期码（年）、出生日期码（月）、出生日期码（日）、顺序码。

7、支持对源数据库的 DDL 进行抽取，包括源数据环境的主外键约束关系、用户自定义、表空间定义、触发器、过程、链接等，自动加载到目标端环境。

4）支持脱敏算法秘钥机制

脱敏算法秘钥机制，即脱敏算法加秘钥生成新脱敏算法，并且秘钥每隔一段时间脱敏管理员可自定义手工修改，修改秘钥后脱敏后的数据和原脱敏算法脱敏后的数据不同，便于脱敏规则安全管理。

5）访问控制机制

★脱敏系统具有完善的访问控制管理机制：

1、具有完善的角色、权限管理体系，实现三权分立（管理员、操作员、审计员）；

2、根据不同的用户对数据源进行权限管理。

6）脱敏结果不可逆

脱敏结果不可通过通过任意方式进行数据还原。敏感数据脱敏的结果不可通过解密算法、原值备份等方式进行敏感数据还原。

7）脱敏作业调度

多样化的脱敏调度，可以灵活的自定义脱敏作业的时间，无需人为干涉。

- 支持脱敏作业立即调度；
- 支持脱敏作业在指定时间调度并只执行一次；
- 支持在一定时间周期内，每天或每隔几天执行一次；
- 支持在一定时间周期内，每周固定几天执行；

★脱敏作业监控动态化，可实时查看脱敏作业的处理效率（如：10000行/s，以及每个作业的进度）、运行时间及状态。通过数据抽样方式，将脱敏前后的数据以报表方式直观展现。(投标时要求提供功能截图)

8）★脱敏系统支持丰富的数据源

1、支持市面上主流的关系型数据库作为脱敏的源和目标，具体包括 Oracle、mysql、ms sqlserver、greenplum、ingre vectorwise、intersystems cache、kingbaseEs、lucaidDb、mariadb、maxdb(sap db)、monetdb、msaccess、native mondrian、neoview、netezza、oracle rdb、postgresql、redshit、sqllite、sybase、terdata、vertica 等。

2、支持 txt、csv、dmp、dbf 等文件类型作为脱敏源和目标，且支持远程 ftp 和 sftp 发现(投标时要求提供功能截图)

3、支持 hdfs，hive，impala 和 kafka 等大数据平台。

4、支持将 Oracle dump 文件作为脱敏源，支持 dump 到 dump，dump 到库的脱敏方式。(投标时要求提供功能截图)

四、实施及服务

4.1 项目验收及交付

1、合同签订后 15 日内确定系统实施、二次开发、项目管理、项目测试/实施的方案，向采购人提供上述文档并经采购人审查通过；

2、合同签订后 2 个月内完成全部系统建设、完成所有相关系统的数据采集，完成相关培训工作，经测试运行，由采购人组织初验合格后进入系统试运行。

3、整体试运行期满 4 个月后，由采购人组织验收，正式验收通过后正式交付使用，并进入免费维护期。

4、项目建设过程中进行的第三方测试费用和项目竣工验收，所需费用由中标人承担（投标人需承诺承担此费用）。

4.2 项目实施要求

本项目建设是一个涉及到区域范围内众多医疗机构的系统工程，不可能一蹴而就，其是一个复杂的过程，是医疗卫生信息化应用和管理从传统模式走向数字化，信息化，智能化，现代化的磨合过程。信息化建设的过程是一个不会结束的过程，它将随着医疗卫生的发展，医学技术和信息技术的发展，医疗卫生管理的发展而不断发展。

本项目建设是基于现有已建业务系统基础上进行的扩展和改造建设项目,投标人应承诺维持现有业务系统不变更、系统数据不中断。

同时为实现本次项目新增业务系统与现有业务系统的无缝衔接

达到项目整体目标，中标人应承诺配合采购人完成业务系统的集成整合工作。

投标厂商应具备较强的行业经验、信息系统集成能力和质量管理体系，同时需要根据项目建设内容和进度需要，**派驻具有一定资质能力水平的成员组成项目小组对医院信息系统进行实施及服务**。投标人应承诺在项目合同签订后的3个月内完成系统调研、培训、数据准备和系统上线等工作。期间可能由于用户的需要及政策的变化而对系统进行相应的客户化修改，投标厂商必须无条件满足。

4.3 培训要求

投标人须在采购人指定的地点提供操作及维护培训，投标人须在投标文件中提供详细的培训计划，包括培训内容、培训时间等。

投标人提供的负责培训的人员应具有相关应用系统开发经验。

技术培训费用应包含在投标总价中。

4.4 售后服务要求

1、投标人必须根据本次招标文件所制定的目标和范围，提出相应的售后服务方案。

2、所有软件产品的质保期自本项目终验合格书签订之日起开始计算。对于应用系统开发，投标人应提供壹年免费维护和承诺永久技术支持，包括各种突发事件采取应急措施等（且免费维护期内必须承诺免费开放接口给第三方系统）。

3、要求投标人承诺项目验收后提供叁年的软件 7*24 小时售后技术服务，30 分钟内做出明确响应和安排，响应时间是接到用户需求电话后 2 小时内到达现场（郊区可延长到 4 小时）。包括免费升级、功能完善、故障排除、性能调优、技术咨询等，并负责系统的开发、集成，处理、协调与各系统软件等供应商的关系。

4、系统维护期间承诺提供固定 1 名技术客服人员随时提供技术支持服务，如人员调整需在人员变动后 7 个工作日内通知滨江区社发局相关负责人。

5、在质保期满时，投标人的工程师和采购人代表对所有产品进行另一次测试，任何故障须由投标人自费解决并取得采购人的认可。

6、中标人在免费维护期满后应向业主提供如质保期内的售后服务，并收取相应费用。

7、质保期内各类维护费用等均由投标人须自行承担。

4.5 知识产权

▲本项目的最终用户为杭州市滨江区社发局，项目的知识产权归采购人所有，系统中涉及中标人此项目投标前已经开发的产品知识产权仍归属于中标人所有。为了项目维护需要中标人必须提供本项目新开发部分的所有源代码和开发文档，采购人有权对软件进行修改。未经同意，中标人不得擅自扩散或提供给第三方使用，但经采购人允许在本系统应用、二次开发或升级除外。中标人对采购人提供的业务资料、技术资料应严格保密，不得扩散。

中标人须对所投产品、方案、技术、服务等拥有合法的占有和处

置权，并对涉及项目的所有内容可能侵权行为指控负责，保证不伤害采购人的利益。在法律范围内，如果出现文字、图片、商标和技术等侵权行为而造成的纠纷和产生的一切费用，采购人概不负责，由此给采购人造成损失的，中标人应承担相应后果，并负责赔偿。

4.6 项目应提交的成果和电子文档

项目应提交的成果和电子文档，包括但不限于：

- 需求分析报告；
- 系统实施方案；
- 系统测试分析报告；
- 安装维护手册；
- 使用操作手册；
- 培训资料；
- 数据结构说明文档。

①产品所有技术性能规格及参数：应符合招标文件和中标方投标文件所要求的技术标准及服务标准。系统运行稳定，无故障，数据无错误。

②验收工作由招标方和中标方共同进行。在验收时，中标方应向招标方提供货物的相关资料，按招标方提出的方式验收。由招标方对货物的质量、规格和数量其他进行检验。如发现质量、规格和数量等任何一项与采购要求规定不符，招标方有权拒绝接受。

5、验收文件的签署：由中标方撰写服务完成报告，由招标方委派的负责人在审核后签署。