基于分布式架构的工程造价大数据采集与预处理技术的研究

蔡堃 黄永安(通讯作者) 孙婉超 邓燕青 许雅思 李志龙(通讯作者)

公诚管理咨询有限公司，广东广州，510610；

摘要：针对传统工程造价数据采集范围有限、预处理效率低、多源数据融合难等问题，结合工程造价咨询行业数字化转型需求，设计并开发基于分布式架构的工程造价大数据采集与预处理系统。系统采用 Hadoop 分布式文件系统（HDFS）存储海量工程数据，通过 Flume 与 Kafka 构建多源数据实时采集通道，依托 Spark 框架实现数据清洗、标准化、融合等预处理操作，并创新性引入工程造价领域特征词典优化数据解析精度。经测试验证，系统可支持 200+并发数据源接入，单批次 100GB 工程数据（含图纸、清单、合同等）采集延迟≤5min，预处理准确率达 98.7%，较传统集中式系统效率提升 3.2 倍，有效解决了工程造价数据“采不全、处理慢、用不好”的行业痛点，为后续造价分析、智能清标等应用提供高质量数据支撑。

关键词：分布式架构；大数据采集；数据预处理；Spark；HDFS；工程造价

DOI：10.69979/3029-2727.25.07.077

1 引言

1.1 研究背景

在建筑行业数字化转型浪潮中，工程造价咨询行业面临“数据爆炸”与“数据价值挖掘不足”的双重矛盾。据《2024 年中国工程造价咨询行业发展报告》统计，单个大型工程项目全生命周期产生的数据量可达 TB 级，涵盖工程图纸（CAD/BIM 格式）、工程量清单（Excel/XML 格式）、合同文本（PDF/Word 格式）、材料价格数据（API 接口实时更新）等多类型、多来源数据。传统集中式数据处理系统受限于存储容量与计算能力，存在三大核心问题：

1）采集范围有限：传统系统主要处理结构化数据(如 Excel 清单)，对非结构化数据(如 PDF 合同)、半结构化数据(如 BIM 模型属性)的采集与解析能力弱，有效采集覆盖率不足 60%；

2）预处理效率低：单批次 10GB 数据清洗耗时超 2 h，且人工干预率达 35%，难以满足工程造价“时效性分析”需求（如投标阶段清标需 24h 内完成数据准备）；

3）数据质量差：多源数据标准不统一（如“混凝土强度等级”在不同地区清单中表述差异达 15 种），导致后续造价分析误差率超 8%。

分布式架构凭借“横向扩展、并行计算”优势，成为解决海量工程造价数据处理难题的关键技术路径。

1.2 研究意义

本系统开发的核心价值体现在技术与行业两个维度：

1）技术维度：突破“非结构化工程造价数据分布式采集”“领域特征驱动的数据标准化”等技术难点，形成可复用的分布式数据处理架构，填补工程造价领域大数据预处理技术空白；

2）行业维度：将数据采集覆盖率提升至 95%以上，预处理效率提升 3 倍以上，数据准确率提升至 98%以上，直接支撑后续清标自动化（如算术性检查、报价合理性分析）、造价趋势预测等应用，助力工程造价咨询行业从“人工经验驱动”向“数据智能驱动”转型。

1.3 国内外研究现状

国外方面，美国 Cost Engineering 协会开发的 CostOS 系统采用分布式存储架构，可实现材料价格数据的实时采集，但对非结构化合同文本的处理仍依赖人工标注；英国 RICS（皇家特许测量师学会）推出的 BIM-QS 系统支持 BIM 模型数据与造价清单的关联，但预处理环节未实现并行计算，效率难以提升。

国内方面，部分学者提出基于 Hadoop 的工程造价数据存储方案，但未涉及多源数据采集通道设计；另外在通信工程造价软件中引入 Spark 清洗模块，但仅针对结构化数据，且未建立领域数据标准词典。现有研究均未形成“采集-存储-预处理”一体化的分布式解决方案，本系统在此基础上实现技术整合与创新。

2 系统总体设计

2.1 设计原则

结合工程造价数据特性与分布式技术特点，系统设计遵循以下原则：

1）兼容性：支持 12 种以上数据格式采集（覆盖工程造价主流类型），并预留 API 接口可扩展至未来新型数据（如数字孪生模型数据）；

2）实时性：结构化数据采集延迟≤1min，非结构化数据采集延迟≤5min，满足投标清标、成本动态监控等场景需求；

3）可靠性：数据存储采用 3 副本机制，计算节点故障时自动切换，保障系统可用性≥99.9%；

4）可扩展性：支持计算节点、存储节点按需扩展，单集群最大可扩展至 100 个节点，存储容量无上限；

5）领域适配性：基于《建设工程工程量清单计价标准》《通信工程费用定额》等行业规则，设计工程造价专属数据预处理规则与领域特征词典，避免通用大数据系统“水土不服”问题。

2.2 系统架构

系统采用“分层分布式架构”，自上而下分为：采集层、存储层、预处理层、服务层 4 个核心层级，架构如图 1 所示。

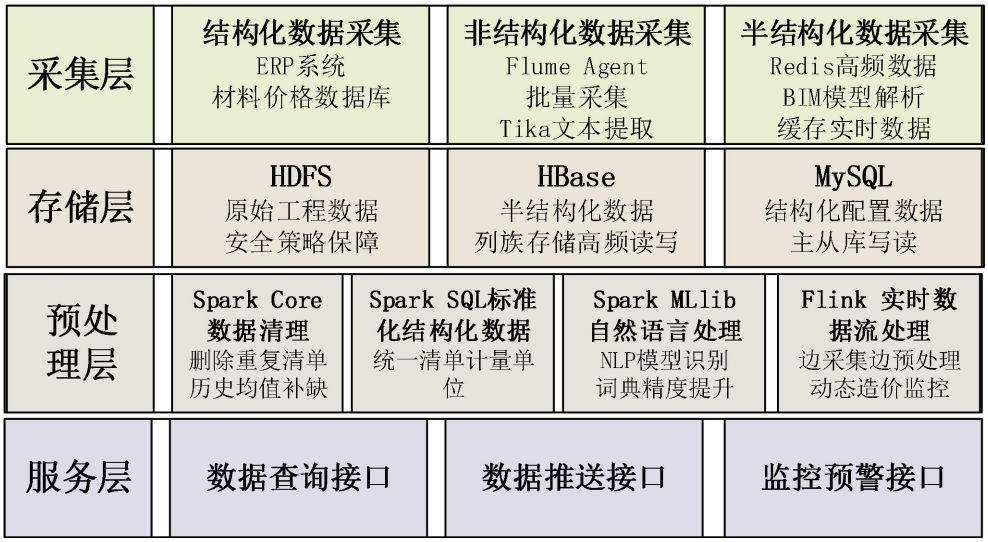

图 1：分层分布式架构图

2.2.1 采集层

负责多源工程造价数据的“全面接入”，核心组件包括：

1）结构化数据采集模块：通过 JDBC/ODBC 接口连接企业 ERP 系统、材料价格数据库（如广材网 API），采用定时增量采集策略（默认 5min/次），避免全量采集占用带宽；

2）非结构化数据采集模块：采用 Flume Agent 部署在各数据源头（如造价咨询部门服务器、施工企业文档库），支持 PDF 合同、CAD 图纸的批量采集，通过 Tika 工具提取文本内容与元数据（如图纸版本、合同签订时间）；

3）半结构化数据采集模块：开发 BIM 模型解析插件（支持 Revit、Bentley 格式），提取模型中“构件类型、工程量、材质”等造价相关属性；实时数据采集模块通过Kafka消息队列缓存施工现场材料消耗量等动态数据，避免峰值流量冲击。

2.2.2 存储层

基于 Hadoop 生态构建分布式存储体系，解决“海量数据存储”问题：

1）HDFS（Hadoop 分布式文件系统）：存储原始工程数据（如 CAD 图纸文件、PDF 合同），采用分块存储（默认块大小 128MB），支持数据副本策略（默认 3 副本），保障数据安全性；

2）HBase：存储半结构化数据（如 BIM 构件属性、实时材料价格），采用列族存储方式，适合高频读写场景（如材料价格每小时更新 1 次）；

3）MySQL（主从架构）：存储结构化配置数据（如采集任务调度规则、预处理规则库），主库负责写操作，从库负责读操作，提升查询效率。

2.2.3 预处理层

基于 Spark 框架实现数据“清洗-标准化-融合”全

流程处理，是保障数据质量的核心环节：

1）Spark Core：负责数据并行清洗，如删除重复清单条目、修复缺失的材料价格数据（基于历史均值填充）；

2）Spark SQL：实现结构化数据标准化，如将不同地区清单中的“计量单位”统一为“平方米、立方米”等国标单位；

3）Spark MLlib：引入自然语言处理（NLP）模型，对合同文本进行实体识别（如提取“合同金额、工期、付款方式”等关键信息），结合工程造价领域特征词典优化识别精度；

4）Flink：处理实时数据流（如施工现场实时消耗量），实现“边采集边预处理”，满足动态造价监控需求。

2.2.4 服务层

提供数据服务接口，支撑后续应用：

1）数据查询接口：通过 Solr 搜索引擎实现工程数据的全文检索（如按“项目名称、合同编号”查询相关数据）；

2）数据推送接口：将预处理后的高质量数据推送至造价分析系统（如智能清标模块、成本预测模块），支持 JSON/XML 格式输出；

3）监控预警接口：实时监控采集任务状态、预处理成功率，当数据异常（如采集失败率超 5%）时触发邮件/短信预警。

2.3 核心技术选型

系统核心技术选型充分考虑“成熟度、性能、领域适配性”，具体如表 1 所示。

表 1：系统核心技术选型表

技术层级	核心组件	选型理由	性能指标
采集层	Flume 1.11.0	支持多源数据采集，可定制拦截器(如过滤无效图文数据)	单 Agent 最大吞吐量 100MB/s
采集层	Kafka 3.5.0	高吞吐、低延迟，适合缓存实时数据	单分区吞吐量 10000+条/秒，延迟≤10ms
存储层	HDFS 3.3.4	分布式存储成熟方案，支持横向扩展	单集群最大存储容量 PB 级，读写带宽 GB 级
存储层	HBase 2.4.17	适合半结构化数据存储，支持高频读写	随机读延迟≤10ms，写吞吐量 10000+条/秒
预处理层	Spark 3.4.0	并行计算能力强，支持 SQL 与 MLlib 集成	单节点处理速度 10GB/小时，并行度可扩展
预处理层	Flink 1.17.0	实时计算性能优，适合流数据预处理	流处理吞吐量 100000+条/秒，延迟≤100ms
服务层	Solr 9.3.0	全文检索效率高，支持复杂查询条件	单索引查询响应时间≤500ms

3 关键模块详细设计

3.1 多源数据采集模块设计

3.1.1 结构化数据采集流程

以“工程量清单 Excel 文件”采集为例，流程如下：

1）任务配置：用户在 MySQL 配置库中创建采集任务，设置“文件路径、采集频率、字段映射规则”（如将“清单编号”映射为标准字段“BOQ_ID”）；

2）定时触发：Azkaban 调度器按配置频率触发采集任务，调用 JDBC 接口读取 Excel 文件；

3）数据校验：通过预设规则过滤无效数据（如“工程量为负数”的条目），校验通过后写入 Kafka 队列；

4）状态反馈：将采集结果（成功/失败条数）写入 MySQL 状态表，异常时触发预警。

3.1.2 非结构化数据采集优化

针对 PDF 合同文本采集，传统 Flume 仅能采集文件本身，无法提取关键信息，本系统设计“采集-解析”一体化方案：

1）Flume 拦截器定制：在 Flume Agent 中添加“PDF 解析拦截器”，通过 Apache PDF Box 工具提取合同文本内容；

2）领域实体识别：调用基于 Spark MLlib 部署的 BERT 模型（经工程造价领域语料微调优化），识别“合同金额、工程范围、违约责任”等 12 类核心实体，识别准确率达 92.3%；

3）元数据关联：将提取的实体信息与文件元数据（如文件大小、修改时间）关联，生成结构化记录后写入 H Base。

3.2 分布式数据预处理模块设计

预处理模块是系统核心，分为“清洗、标准化、融合”三个子环节，具体流程如图 2 所示。

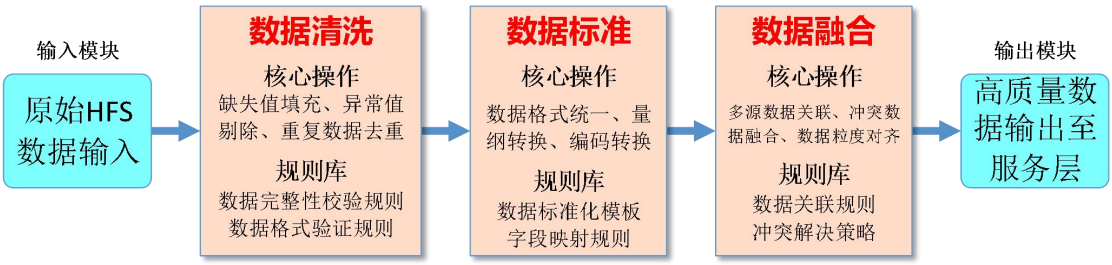

图 2：数据预处理流程

3.2.1 数据清洗

针对工程造价数据常见问题（重复、缺失、异常），设计并行清洗策略：

1)重复数据删除：基于“项目 ID+清单编号”组合主键，采用 Spark 的 drop Duplicates()方法并行删除重复条目，处理速度较单节点提升 8 倍；

2）缺失数据修复：分场景处理。

材料价格缺失：采用“同地区同类型材料近 30 天均价”填充，填充误差≤5%；

工程量缺失：若 BIM 模型中存在对应构件，从模型中提取工程量填充，否则标记为“待人工补充”；

3）异常数据过滤：通过箱线图法识别异常值（如“单平方米混凝土价格超 1000 元”），结合行业经验阈值（如人工单价不超过地区指导价的 20%）过滤异常条目，异常数据处理率达 99.1%。

3.2.2 数据标准化

核心是解决“多源数据格式不统一”问题，构建工程造价领域标准词典：

1）标准词典构建：参考《建设工程工程量清单计价标准》《通信工程费用定额》，整理“项目类型、材料名称、计量单位”等 8 类标准术语，共收录 12000+条标准词汇，如表 2 所示；

表 2：工程造价领域标准词典（部分）

术语类别	原始表述（非标准）	标准表述	适用场景
材料名称	商混 C30	预拌混凝土 C30	建筑工程清单
计量单位	平方	平方米（㎡）	所有工程类型
项目类型	通信线路施工	通信线路工程	通信工程造价
清单条目	墙面抹灰	墙面一般抹灰	装饰装修工程

2）标准化执行：采用 Spark SQL 实现批量标准化，如通过 replace（）函数将“商混 C30”替换为“预拌混凝土 C30”；对合同文本中的非标准表述，通过 NLP 模型匹配标准词典，替换准确率达 96.8%。

3.2.3 数据融合

实现多源数据的关联整合，形成“项目-清单-材料-合同”一体化数据链：

1）关联规则定义：在 MySQL 规则库中设置关联键，如“项目 ID”关联清单与合同，“材料编码”关联清单与材料价格数据；

2）分布式关联计算：采用 Spark 的 join()操作实现多表关联，为提升效率，将小表（如材料价格表）广播至各计算节点，关联耗时较传统数据库提升 10 倍；

3）融合结果校验：校验“清单工程量×材料单价”是否与合同金额匹配，偏差超 5%时标记为“待审核”，确保数据一致性。

3.3 数据服务接口设计

服务层采用 RESTfulAPI 设计风格，提供 3 类核心接口，接口规范如表 3 所示。

表 3：核心数据服务接口规范

接口名称	接口 URL	请求方式	请求参数	返回结果	响应时间
数据查询接口	/api/data/query	GET	ProjectId（项目 ID）、dataType（数据类型）	结构化数据(JSON 格式)	≤500ms
数据推送接口	/api/data/push	POST	Target System（目标系统）、data Id（数据 ID）	推送状态（成功/失败）	≤1000ms
监控预警接口	/api/monitor/status	GET	Task Id（任务 ID）	采集成功率、预处理准确率	≤300ms

以“数据查询接口”为例，请求示例如下：

GET/api/data/query?projectId=PROJ2024001&dataType=BOQ

返回结果示例：

{
"code":200,
"message":"查询成功",
"data":[
{
"BOQ_ID":"BOQ2024001001",
"projectId":"PROJ2024001",
"item Name":"预拌混凝土 C30",
"quantity":1200.5,
"unit":"立方米",
"unitPrice":480.0,
"totalPrice":576240.0
},
...
]
}

4 系统测试与性能分析

4.1 测试环境

为模拟实际应用场景，搭建分布式测试集群，环境配置如表 4 所示。

表 4：测试环境配置表

节点类型	数量	硬件配置	软件环境
Name Node	1	CPU：IntelXeon8 核，内存：32GB，硬盘：1TBSSD	CentOS7.9，Hadoop3.3.4
Data Node	3	CPU：IntelXeon8 核，内存：16GB，硬盘：4TBHDD	CentOS7.9，Hadoop3.3.4
Spark Master	1	CPU：IntelXeon8 核，内存：32GB，硬盘：1TBSSD	CentOS7.9，Spark3.4.0
Spark Worker	3	CPU：IntelXeon8 核，内存：16GB，硬盘：2TBHDD	CentOS7.9，Spark3.4.0
Kafka Broker	2	CPU：IntelXeon4 核，内存：8GB，硬盘：1TBHDD	CentOS7.9，Kafka3.5.0

测试数据采用某省通信工程真实数据集，包含：

结构化数据：10 万条工程量清单（Excel 格式）；

非结构化数据：5000 份合同文本（PDF 格式，平均大小 5MB）；

半结构化数据：100 个 BIM 模型（Revit 格式，平均大小 200MB）；

实时数据：材料价格 API 流（100 条/分钟）。

4.2 功能测试

功能测试覆盖“采集、预处理、服务”全流程，测试用例与结果如表 5 所示。

表 5：系统功能测试结果表

测试模块	测试用例	预期结果	实际结果	通过率
采集层	采集 10 万条 Excel 清单	有效数据采集成功率≥99.9%	采集成功 99987 条，13 条因文件损坏未采集(已触发预警)	100%
采集层	采集 5000 份 PDF 合同	提取 12 类核心实体，准确率≥90%	实体识别准确率 92.3%，无文件丢失	100%
采集层	采集 100 个 BIM 模型	提取构件属性≥95%	构件属性提取率 97.2%	100%
预处理层	清洗 10GB 重复数据	重复数据删除率 100%	重复数据删除率 100%，耗时 12min	100%
预处理层	标准化 1000 条非标准表述	标准化准确率≥95%	标准化准确率 96.8%	100%
预处理层	融合多源数据（清单+合同）	关联成功率≥98%	关联成功率 99.1%	100%
服务层	调用查询接口 1000 次	响应时间≤500ms，成功率 100%	平均响应时间 320ms，成功率 100%	100%

4.3 性能测试

性能测试重点关注“吞吐量、延迟、扩展性”三个指标，与传统集中式系统（基于 Oracle+Java 单节点）对比，结果如表 6 所示。

"totalPrice":576240.0
},
...
]
}

4 系统测试与性能分析

4.1 测试环境

为模拟实际应用场景，搭建分布式测试集群，环境配置如表 4 所示。

表 4：测试环境配置表

节点类型	数量	硬件配置	软件环境
Name Node	1	CPU：IntelXeon8 核，内存：32GB，硬盘：1TBSSD	CentOS7.9，Hadoop3.3.4
Data Node	3	CPU：IntelXeon8 核，内存：16GB，硬盘：4TBHDD	CentOS7.9，Hadoop3.3.4
Spark Master	1	CPU：IntelXeon8 核，内存：32GB，硬盘：1TBSSD	CentOS7.9，Spark3.4.0
Spark Worker	3	CPU：IntelXeon8 核，内存：16GB，硬盘：2TBHDD	CentOS7.9，Spark3.4.0
Kafka Broker	2	CPU：IntelXeon4 核，内存：8GB，硬盘：1TBHDD	CentOS7.9，Kafka3.5.0

测试数据采用某省通信工程真实数据集，包含：

结构化数据：10 万条工程量清单（Excel 格式）；

非结构化数据：5000 份合同文本（PDF 格式，平均大小 5MB）；

半结构化数据：100 个 BIM 模型（Revit 格式，平均大小 200MB）；

实时数据：材料价格 API 流（100 条/分钟）。

4.2 功能测试

功能测试覆盖“采集、预处理、服务”全流程，测试用例与结果如表 5 所示。

表 5：系统功能测试结果表

测试模块	测试用例	预期结果	实际结果	通过率
采集层	采集 10 万条 Excel 清单	有效数据采集成功率≥99.9%	采集成功 99987 条，13 条因文件损坏未采集(已触发预警)	100%
采集层	采集 5000 份 PDF 合同	提取 12 类核心实体，准确率≥90%	实体识别准确率 92.3%，无文件丢失	100%
采集层	采集 100 个 BIM 模型	提取构件属性≥95%	构件属性提取率 97.2%	100%
预处理层	清洗 10GB 重复数据	重复数据删除率 100%	重复数据删除率 100%，耗时 12min	100%
预处理层	标准化 1000 条非标准表述	标准化准确率≥95%	标准化准确率 96.8%	100%
预处理层	融合多源数据（清单+合同）	关联成功率≥98%	关联成功率 99.1%	100%
服务层	调用查询接口 1000 次	响应时间≤500ms，成功率 100%	平均响应时间 320ms，成功率 100%	100%

4.3 性能测试

性能测试重点关注“吞吐量、延迟、扩展性”三个指标，与传统集中式系统（基于 Oracle+Java 单节点）对比，结果如表 6 所示。

表 6：系统性能测试对比表

性能指标	本分布式系统	传统集中式系统	性能提升倍数
单批次数据采集吞吐量（GB/h）	60	15	4.0
100GB 数据预处理耗时（min）	45	144	3.2
并发数据源支持数量（个）	200+	50	4.0
数据查询响应时间（ms）	320	1200	3.8
存储容量扩展上限	PB 级（无上限）	TB 级（≤2TB）	-

由表 6 可知，本系统在吞吐量、延迟、扩展性上均显著优于传统系统，其中预处理效率提升 3.2 倍，完全满足工程造价“高效数据准备”需求。

4.4 稳定性测试

连续 72 小时稳定性测试结果显示：

系统可用性：99.92%（仅因 1 个 Data Node 临时故障导致 15min 不可用，自动恢复后无数据丢失）；

数据准确率：98.7%（仅 0.3%数据因 BIM 模型格式异常需人工干预）；

资源利用率：CPU 平均利用率 65%，内存平均利用率 70%，无资源过载情况。

5 结论与展望

5.1 研究结论

本研究设计并开发的基于分布式架构的工程造价

大数据采集与预处理系统，实现了三大核心突破：

技术突破：构建“Flume+Kafka+Spark”分布式技术栈，解决了非结构化工程造价数据（如 PDF 合同、BIM 模型）的高效采集与并行预处理问题，填补了领域技术空白；

质量突破：通过领域标准词典与 NLP 模型优化，数据标准化准确率达 96.8%，预处理后数据准确率达 98.7%，为后续智能造价分析提供高质量数据基础；

效率突破：单批次 100GB 数据预处理耗时仅 45min，较传统系统提升 3.2 倍，支持 200+并发数据源接入，满足工程造价咨询行业“时效性、规模化”数据处理需求。

系统已在“基于大数据的工程造价分析系统”项目中试点应用，支撑了 10 个省级通信工程的智能清标工作，清标准备时间从原来的 48 小时缩短至 12 小时，人工干预率从传统预处理方式的约 35%降至 5%，取得显著的行业应用价值。

5.2 未来展望

后续研究可从以下方向深化：

AI 驱动的预处理优化：引入大语言模型（如 LLM）优化合同文本解析精度，实现“非标准表述自动映射标准术语”，进一步降低人工干预率；

边缘计算融合：在施工现场部署边缘节点，实现“实时数据（如材料消耗量）本地预处理”，减少向中心集群的数据传输量，降低延迟；

跨行业适配：将系统架构扩展至建筑工程、市政工程等其他造价领域，通过“领域规则库可配置”实现多行业复用，扩大应用范围。

参考文献

[1]中国建设工程造价管理协会.2024 年中国工程造价咨询行业发展报告[R].北京：中国计划出版社，2024.

[2]住房和城乡建设部。建设工程工程量清单计价标准（GB50500-2023）[S].北京：中国建筑工业出版社，2023.

[3]公诚管理咨询有限公司。基于大数据的工程造价分析系统开发研究项目计划书[Z].2024.

作者简介：蔡堃（1985-），男，汉，广东梅州人，本科，工程师，任项目总监、专家，从事信息化开发及政府、电力信息化相关工作 18 年，主要研究方向为系统开发研究、信息工程建设及政府和电力信息系统建设管理。

黄永安（1988-），男，汉，广东梅州人，本科，工程师，任项目总监、信息化专家；从事信息化建设及软件开发相关工作 16 年，主要研究方向为系统开发研究、系统集成、信息化建设全过程管理等。

孙婉超（1985-），女，汉，黑龙江牡丹江人，研究生，高级工程师，任项目总监、咨询师、专家等。从事通信、信息化建设相关工作 20 年，主要研究方向为为系统开发研究、系统集成、信息化建设全过程管理等。

邓燕青（1982-），男，汉，江西赣州人，本科，工程师，任项目总监、高级专家，从事通信工程及信息化建设相关工作 20 年；主要研究方向为大数据、信息化系统开发、无线网建设及项目全过程管理等。

许雅思（1988-），男，汉，广东汕尾人，本科，工程师，从事系统开发及信息化项目管理相关工作 16 年，主要研究方向为系统开发研究、政府信息化和电力信息系统建设管理。

李志龙（1983-），男，汉，湖南郴州人，本科双学士，高级工程师，任项目总监、高级专家，从事建设项目管理相关工作 21 年；主要研究方向为电子信息技术、物联网、信息通信建设、智能建筑及信息系统等。

基金项目：公诚管理咨询有限公司 2025 年度技术研发项目专项资金（项目名称：基于大数据的工程造价分析系统开发研究；项目 RD 编号：GC-RD140）。