长沙农村商业银行股份有限公司

内容管理平台项目需求

信息技术部

2020年11月

目 录

[一、项目背景 1](#_Toc55564962)

[1.1. 存量影像数据庞大，且仍在持续上涨 1](#_Toc55564963)

[1.2. 影像资料储存方式各异，存在重大安全隐患 2](#_Toc55564964)

[1.3. 缺乏统一管理，无法挖掘影像数据价值 2](#_Toc55564965)

[**二、建设目标** 3](#_Toc55564966)

[**三、建设原则** 4](#_Toc55564967)

[**四、建设内容** 5](#_Toc55564968)

[4.1 系统架构设计 5](#_Toc55564969)

[4.2 系统功能需求 7](#_Toc55564970)

[4.2.1影像采集 7](#_Toc55564971)

[4.2.2影像传输 7](#_Toc55564972)

[4.1.3影像调阅 7](#_Toc55564973)

[4.1.4影像批注 8](#_Toc55564974)

[4.1.5影像处理 8](#_Toc55564975)

[4.1.6影像压缩 8](#_Toc55564976)

[4.1.7影像存储 9](#_Toc55564977)

[4.1.8分页索引 9](#_Toc55564978)

[4.1.9数据安全 9](#_Toc55564979)

[4.1.10识别功能 10](#_Toc55564980)

[4.1.11系统监控 11](#_Toc55564981)

[4.3 分布式存储需求 11](#_Toc55564982)

[4.4 系统部署方式 12](#_Toc55564983)

# 一、项目背景

随着我行柜面、会计、信贷等业务量的不断攀升及无纸化办公的推行，大量的业务资料、凭证等影像资料需要提交、审核和保存，传统的手工处理方式速度慢、效率低、存储难、费用高，已经不适应当前业务发展的需求。

为加速推进我行数字化战略转型，推动业务处理电子化、自动化方向发展，加强我行影像等非结构化数据统一管理，规避影像资料泄露等业务风险，我部需统一规划并建设内容管理平台。参考同行业成功案例，并结合本行实际情况，需要对业务凭证、合同等资料进行数字化处理，建设专业的数据内容管理平台，为全行非结构化数据的管理保驾护航。

## 存量影像数据庞大，且仍在持续上涨

我行自主建设的业务系统，如事后监督系统、理财录音录像（双录）系统、信用卡影像信息平台等，每天都会产生大量的业务影像数据。现已有存量数据约7TB，并以每天约20GB的增长量在飞速上涨。具体情况如下所示。

表1：当前系统数据量

| **系统名称** | **存量数据** | **增量数据（/天）** |
| --- | --- | --- |
| 事后监督系统 | 4T | 15G |
| 信用卡影像平台 | 1.3T | 2G |
| OA协同办公系统 | 1T | 2G |
| 录音录像（双录）系统 | 370G | 1G |
| 放款业务影像 | 15G | 100MB |
| Q房贷 | 10G | 200MB |
| 合计 | 6.7T | 20.3G |

## 影像资料储存方式各异，存在重大安全隐患

我行自主建设的部分业务系统在数据采集、存储等环节，存在着影像数据泄露、丢失等风险。如客户经理用手机拍摄的客户贷款资料等，很容易因手机丢失或通过网络方式传播，而导致客户数据的泄露。又比如部分业务系统的影像数据会暂存在客户端本地进行缓存，或只存放在应用服务器本地磁盘而不做数据多副本备份，都会因硬件发生故障等情况导致影像数据的丢失。

## 缺乏统一管理，无法挖掘影像数据价值

目前我行建设的各个系统采用的是传统的“烟囱式”架构，在影像的采集、传输、存储、展示等过程都是“各自为政”，影像数据缺乏统一的有效管理。各自系统成为了信息孤岛，不利于全行的业务发展及信息系统建设，也无法深入挖掘影像数据的潜在价值。

# **二、建设目标**

内容平台的建设旨在构建全行级的内容管理基础架构，制定非结构化数据管理的统一标准，实现各种海量的非结构化数据的统一安全的管理、高效便捷的访问、并实现信息的全面整合和共享，最终建立全行内容的统一视图，完成非结构化数据规划治理。

**1、数据资源整合**

对我行现有的非结构化数据资源实现集中、整合，构建全行级的内容管理基础架构。优化现有非结构数据的传输、存储等流程，制定非结构化数据管理的统一标准，实现各种海量的非结构化数据的统一安全的管理、高效便捷的访问、并实现信息的全面整合和共享，避免传统的“烟囱式”系统建设导致影像模块重复建设，降低经营成本。

**2、全过程管理**

实现非结构化数据从采集、管理到利用、归档的整个生命周期管理，提供标准化的开发工具、服务接口和工具插件，支持与各业务系统的集成及二次开发。

**3、支撑业务系统**

为行内各类应用系统提供统一的影像数据采集、影像调阅、影像展现、影像传输、影像存储等功能，对全行影像数据的集中管理和存储，提供统一的影像支持服务。实现事后监督系统、录音录像（双录）系统、信用卡影像信息平台、放款业务影像等业务系统的集成接入。支撑后续行内业务系统的建设，支撑大数据平台及其业务应用的数据交互。

# **三、建设原则**

内容管理平台在建设、设计和系统实现过程中，应遵循如下指导原则：

**1、经济性：**根据我行科技发展规划，以及基于我行现有业务系统分析，对三年内的数据量及业务进行合理评估，确定内容管理平台存储规模，后续根据实际情况再逐步优化扩容。

**2、可扩展性：**架构设计与功能划分模块化，考虑各接口的开放性、可扩展性，便于系统的快速扩展与维护，便于第三方系统的快速接入。

**3、可靠性：**系统采用的系统结构、技术措施、开发手段都应建立在已经相当成熟的应用基础上，在技术服务和维护响应上同用户积极配合，确保系统的可靠；对数据指标要保证完整性，准确性。

**4、安全性：**针对系统级、应用级、网络级，均提供合理的安全手段和措施，为系统提供全方位的安全实施方案，确保企业内部信息的安全。

**5、先进性：**借鉴业内内容管理平台建设的实践，可采用微服务应用+分布式存储使平台具有良好的先进性和弹性，支撑当前及未来非结构化数据应用需求。

**6、平台性：**归纳整理各业务系统影像应用需求，利用内容管理平台的影像能力，为业务系统的建设提供平台。

# **四、建设内容**

本次项目内容范围包括分布式数据库存储软件 （巨杉SequoiaDB）、内容管理平台应用软件实施及售后维护等。

## 4.1 系统架构设计

根据我行IT架构规划，内容管理平台将作为影像应用的基础架构，为全行的应用系统提供支撑，整体架构规划如下：



图1 整体架构规划

1、基础设施组件

基础设施包括影像处理组件和内容管理组件。影像处理组件主要包含影像采集、加工、识别、传输、安全等。内容管理组件主要包含内容存取、目录服务、备份服务等。

2、数据存储

本次平台建设，拟采用分布式数据库存储经业务系统采集的非结构化数据。分布式数据库具有文件自动分片存储、自动合并调阅、多副本保存等非结构化数据存取优势，具备弹性扩展、高并发和高可用性等特性，非常适合内容管理平台海量的非结构化数据存取。

## 4.2 系统功能需求

### 4.2.1影像采集

1、影像采集支持主流各种档次的TWAIN接口的高速扫描仪、高拍仪和平板扫描仪、摄像头等；

2、文件导入：支持TIF、JPG、PDF等格式影像文件批量导入，支持文本、音视频文件等其他非影像文件批量导入。

### 4.2.2影像传输

1、影像传输应具有安全性、可靠性；

2、支持单笔传输、批量传输；

3、支持http、socket等主流通讯协议；

4、支持断点续传、并发、流量控制；

5、支持同步、异步方式传输。

### 4.1.3影像调阅

1、根据用户权限确定用户可查看影像资料内容；

2、单幅影像查看：系统需提供整体缩放、局部放大、旋转功能；

3、支持资料目录树的展示方式，目录树支持右键菜单；

4、可以根据用户权限控制影像文件的下载、打印，支持按目录树、按列表选择打印；

5、支持缩略图显示；

6、支持影像多页显示模式；

7、格式转换：影像、文本等转换成pdf、pdf文件拆分。

### 4.1.4影像批注

1、支持对影像区域进行文字批注、透明框、高亮区域等多种批注方式；

2、支持在同一图像上添加多个批注；

3、支持个性化批注设置，如设置颜色、字体等；

4、批注独立于影像单独保存。

### 4.1.5影像处理

1、影像缩放：影像放大、影像缩小、局部放大；

2、影像旋转：左旋、右旋、翻转；

3、显示尺寸：自适应、适应宽、适应高、原始大小；

4、影像拖动：在显示区域中拖动图像；

5、图像纠偏：纠正倾斜图像。

### 4.1.6影像压缩

支持影像的压缩，可设置压缩比。扫描完成后，需要通过影像压缩技术使整个影像文件降低到理想水平，降低存储的开销和网络传输的成本，提高效率。

### 4.1.7影像存储

1、影像存储：实现对影像及索引数据的存储功能；

2、支持各类非结构化数据文件的存储，例如jpg、png、bmp等图片格式，pdf、word、excel、txt等文档格式；

3、支持文档检入检出，通过将某一文档检出，可以防止不同人同时编辑修改同一文档，从而保证数据的一致性；

4、版本管理：保存同一文件的多个不同版本。

### 4.1.8分页索引

对合同类等多页的影像文件的采集、存储、调阅等，需支持以下功能：

1、多页文件自动分页采集、存储；

2、自动识别、提取每页关键字作为检索索引；

3、支持整本文件调阅；

4、支持按每页关键字索引调阅。

### 4.1.9数据安全

多角度维护数据的安全：

影像采集后在业务流转过程中，需对传递中的影像文件进行加密处理，达到影像的安全传输及处理的要求。

影像调阅及导出时，需对用户调阅的影像文件进行自动化水印处理，确保查阅、导出文件的痕迹可溯源。

支持通过对用户角色、权限、组织等多维度的权限安全管理，控制不同用户对不同数据的操作权限。

### 4.1.10识别功能

1、版面识别

版面识别软件包能够自动从影像数据中提取图像特征信息，然后与事先定义的模板图像信息进行自动匹配，同时快速准确识别几百种图像版面类型。版面识别功能有助于对影像数据进行初步分类归档以及方便后继的OCR识别等处理过程，版面识别软件开发包主要包括影像特征信息自动提取、影像特征信息自动匹配等功能，可识别的版面包括有框线版面和无框线版面。

对于新增票据，要可以通过参数化配置增加版面。

2、OCR识别

OCR识别软件包能够自动识别影像中的数据信息，具有快速和高度智能化的优点。

本模块包括以下主要功能：打印体数字识别；手写体数字识别；打印体字母识别；打印体汉字识别；手写体汉字识别；磁码数字识别。

具有以下特点：高识别率；识别速度快；具有字体学习扩充功能。

3、条码识别

条码识别软件包能实现在图像的指定区域或整幅图像上自动搜索并识别条码，具有方便、准确、快速和高度智能化的优点。

### 4.1.11系统监控

1、硬件监控：针对CPU、内存、存储等基础硬件进行监控；

2、服务监控：对应用服务、数据库服务等进行监控和预警；

3、服务启停：服务启动和停止；

4、自动化测试：测试后台服务运行是否正常。

## 4.3 分布式存储需求

1、分布式架构，支持share-nothing架构，无中心主控节点，以避免单点故障。

2、支持分布式数据存储，支持分库分表。

3、数据分布、分区方式至少支持哈希散列和范围分区两种方式。

4、原生支持多副本高可用机制。不需要通过第三方组件就可以实现高可用。集群中任意一个节点失效都不会影响集群服务能力。失效节点恢复以后可以自动加入集群重新提供服务。

5、原生支持同城灾备。不需要借助第三方组件即可实现集群跨机房（同城的主机房和灾备机房）部署，并且实现副本之间的数据同步。而且任意机房的整体损毁不会造成任何数据丢失。数据副本所在物理位置可以固定，不可自动迁移。

6、数据副本同步方式支持随需配置，在数据表级别配置的强一致性和最终一致性。而且最终一致性的副本同步成功个数可配置。

7、易于维护，能够提供信息收集、性能监控、数据一致性监测等工具。

8、增加数据存储节点即可提升存储空间与计算能力。集群扩容时可以重组集群内数据以适应新的节点。

9、原生支持结构化、半结构化（JSON格式）和非结构化数据的统一存储，即多模式数据库。

10、支持通过API方式访问结构化数据和非结构化数据。

## 4.4 系统部署方式

根据我行的分行网点的地理位置分布区域大这个特点，为保证系统的数据传输效率、稳定性，内容管理平台支持分布式部署。内容管理平台将在支行网点端（由业务系统调用平台组件）部署影像采集、影像传输、查询统计等服务。在中心业务处理端部署影像传输、审批、查询统计功能。在后台管理端部署权限管理、系统管理、影像加工等功能。部署方式如下：



图2 部署方式

在总行部署设备包括：数据库服务器，应用服务器，多台分布式数据库服务器节点。系统可以支持虚拟化部署。在内网和外网之间设置一隔离区，在隔离区服务器中部属前置服务，对于从外网上传的影像，影像文件通过隔离区应用服务，先从外网传到隔离区，防止系统和外网直接通信，保证内网网络安全。