CAsDC-ROADMAP-2017-2020

中国天文数据中心数据管理系统技术路线图

v0.1

中國虛擬天文臺

版本历史

版本	日期	负责人	备注
v0.0	2017-06-22	何勃亮	开始
v0.1	2017-07-07	何勃亮	第一个小阶段版本

更新时间：2017 年 7 月 18 日

目录

插图

第 1 章　前言

本文的目标是简要说明构建满足中国天文数据中心与数据相关的管理与运行需要的一系列基础设施的技术路线图（2017-2020），目标是构建一个综合性的天文数据管理系统，技术路线是从两个方面入手：一是构建数据的元数据体系，二是构建朱雀分布式数据管理系统。两项工作将并列进行，并在底层实现一致性统一。

本文分三个部分进行介绍：

一、目标与路线，提出未来 2-3 年的技术目标，并为此设定了一个技术路线图以及分三期的研发计划。

二、简单介绍了元数据的体系设定。

三、从粗的层次描绘了朱雀系统的技术方案。

第 2 章　目标与路线

一切工作的基础都是围绕数据展开，数据类型主要分为**科学数据**和**运行数据**。

2.1　目标

目标是构建一个综合性的天文数据管理系统，涵盖以下主要功能及目标：

- 基本符合国际虚拟天文台联盟相关协议的统一的元数据规范体系；
- 高性能数据存储引擎；
- 面向用户的易用性管理界面；
- 分布式网络架构，真正形成逻辑上统一、物理上分散的数据存储管理模式；
- 接口化；
- 与中国虚拟天文台云系统无缝对接；
- 模块化设计，对不同对应用用户可以进行定制；
- 应该成为天文大数据的应用引擎，大数据计算服务应该可以对存储在系统中的数据直接读写和分析。

2.2　预期典型应用场景

基于上述目标，可以描绘最终的基本应用场景：

2.2.1　主数据中心

主数据中心作为主控的中心，可以看到所有资源的详细信息，可以通过 Web 界面、手机客户端等了解：

- 数据中心总览；
- 最新和历史的数据归档情况；
- 查看每个数据集的详细信息（背景情况、覆盖图、历史归档、数据共享情况等）；
- 关联异地数据节点（数据中心）的资源情况、网络链接情况；

- 分配节点、数据集的中国虚拟天文台唯一编号；
- 分布式系统管理与运维；
- 筛选、部署数据到中国虚拟天文台数据发布服务器上，供数据发布系统使用，比如数据库、数据文件等；
- 查看和下载每个数据；
- 打包数据；
- 备份数据、数据历史备份情况；
- 用户管理、授权；
- 云接口；
- 其他。

2.2.2 子数据中心或望远镜运行中心

子数据中心，可以是区域数据中心，也可以是某个团组的数据中心，也可以是一个望远镜的运行中心，管理一个或多个特定望远镜的数据集：

- 数据中心总览；
- 最新和历史的数据归档情况；
- 查看和管理等每个数据集的详细信息（背景情况、覆盖图、历史归档、数据共享情况等）；
- 与主数据中心网络链接情况；
- 查看和下载数据；
- 打包数据；
- 备份数据；备份到主数据中心或云中；
- 子用户管理、授权；
- 云接口；
- 其他。

2.2.3 个人用户

个人用户主要可以使用系统的单机版来管理本人的数据，是一个个人数据管理平台：

- 数据总览；
- 最新和历史的数据归档情况；

- 自定义数据集，录入管理数据集的详细信息；
- 与主数据中心网络链接情况；
- 查看和下载数据（可视化）；
- 打包数据；
- 备份数据；备份到主数据中心或云中；
- 云接口；
- 其他。

2.2.4　定制用户

定制用户主要面向特殊需求用户进行软件的定制开发，上述的这些功能都是模块化设计、可以进行裁剪。比如对兴隆观测基地等定制的数据管理系统。

2.2.5　虚拟天文台程序

由于本系统是一个偏向底层的分布式数据管理系统，更侧重于数据管理，因此对虚拟天文台协议接口的包装不多，虚拟天文台协议接口主要由中国虚拟天文台的数据访问接口实现。本系统的各个版本的程序仅提供统一的 VOSpace 数据访问接口。

2.2.6　手机端

手机端主要是基于 HTML5 开发，根据登陆用户的权限，可以查看数据、数据中心的基本情况。

2.3　路线图与工作计划

为达到最终的效果与目标，将有大量的工作需要做。在 2017 年 2-6 月，基于大量的技术细节做了试验和验证，并从基本形成了一个完整的技术方案。整个研发计划在 2017 年下半年全面展开。

总的计划以近期、中期和长期分为三个里程碑进行计划的编制和实施。初步拟定的技术方案可以查看第3章和第4章。

在研发过程中，将以敏捷开发的模式编写开发文档；

2.3.1 近期目标（2017.12）

近期目标以2017年底为近期目标的时间节点，将达到的最终效果如下：

1. 完成元数据库的设计与实现，开发一套元数据管理系统，实现对元数据的管理，并逐步录入望远镜、数据集等的元数据；
2. 完成分布式数据管理系统的一个 `v0.1` 系统，可以实现下述功能：
   (a) 一个基于Web的管理界面；
   (b) 实现对至少三个数据集数据的在线管理。
   (c) 实现的基本功能有：数据元信息提取、数据输入输出、数据备份传输自动化、界面展示等。
   (d) 系统初具雏形。

需要编制的文档有：

- 《中国天文数据中心元数据库规范-v1.0》；
- 《中国天文数据中心分布式存储系统-朱雀-技术手册-v0.1》；

2.3.2 中期目标（2018.12）

中期目标以2018年底为限，目标版本为 `v1.0`，将实现最终场景计划中的绝大部分功能。并且在至少三个异地站点进行部署，并且最终形成一个初步的中国天文数据管理网络。

2.3.3 长期目标（2020.12）

长期目标以2020年底为限，目标版本为 `v2.0`，完全实现最终场景计划中的全部功能，真正实现物理上分散、逻辑上统一，面向不同层次用户、面向应用的中国虚拟天文台数据中心。

第 3 章　元数据

元数据定义和记录关于数据相关的一切可用信息。详细设计参考：《中国天文数据中心元数据库规范-v1.0》。

数据的元数据主要存储在数据库中，也有部分存储在文件系统中。

元数据是分层次结构的，由以下几个部分和类型组成：

- 基础信息元数据
  - 站址
  - 望远镜
  - 观测终端
  - 观测计划
- 科学数据元数据
  - 数据集
  - 专业库（数据集组）
  - 数据表
  - 统计信息
- 观测元数据
  - 日志
  - 观测历史
  - 归档历史
  - 观测计划
- 服务元数据
  - 论文库
  - 项目库
  - 新闻库

3.1 数据元数据

数据元数据的规范重点参考《IVOA Dataset Metadata Model》的规范进行实现。

3.1.1 望远镜元数据

望远镜元数据包括望远镜站点的信息、图片；望远镜的口径、终端、图片、波段、历史等信息。这些信息最终都可以进行可视化的展示。

望远镜的元数据主要描述望远镜相关的站址以及硬件信息。

表 3-1 望远镜元数据

字段	定义	字典	说明
ID	INT8	0-32767	编号
SiteID	INT8	0-32767	站址编号
Name	VARCHAR(100)		中文名称
EnName	VARCHAR(100)		英文名称
Abbr	VARCHAR(20)		缩写
Descrption	TEXT		
Latitude	FLOAT4		望远镜经度
Longitude	FLOAT4		望远镜纬度
Altitude	FLOAT4		望远镜高度

表 3-2 望远镜终端元数据

字段	定义	字典	说明
ID	INT8	0-32767	编号
TelID	INT8	0-32767	望远镜编号
Name	VARCHAR(100)		中文名称
EnName	VARCHAR(100)		英文名称
Descrption	TEXT		

表 3-3　站址元数据

字段	定义	字典	说明
ID	INT8	0-32767	编号
Name	VARCHAR(100)		中文名称
EnName	VARCHAR(100)		英文名称
Abbr	VARCHAR(20)		缩写
Descrption	TEXT		
Latitude	FLOAT4		站点经度
Longitude	FLOAT4		站点纬度
Altitude	FLOAT4		站点高度
Images	[]TEXT	URL	站点图片
Contact	VARCHAR(48)		联系人
ContactTel	VARCHAR(48)		联系人
ContactEMail	VARCHAR(48)		联系人
ContactFax	VARCHAR(48)		联系人

3.1.2　数据集元数据

数据集的元数据记录数据集相关的所有信息：标识、说明、分类、版本、覆盖图等信息。

3.1.3　专题库元数据

专题库是一系列拥有共同属性或标签的数据集的合集。

专题库是一组数据集的合集。一个观测项目也可以称为一个专题库。

3.1.4　数据表元数据

记录每个数据集下每个表格（比如星表）的元数据。参考 IVOA 相关标准制定。

3.1.5　文件元数据

文件元数据分为两层，一层为基本信息元数据，一层为可选的详细信息元数据，比如 FITS 文件的头信息等。

表 3-4 数据集元数据

字段	定义	字典	说明
ID	INT8	0-32767	编号
Name	VARCHAR(100)		中文名称
EnName	VARCHAR(100)		英文名称
Doi	VARCHAR(100)		标识符
Abbr	VARCHAR(20)		缩写
Descrption	TEXT		
Version	VARCHAR(10)		版本
Keywords	VARCHAR(40)	不少于 3 个	关键字
Class	VARCHAR(40)	参看分类表	分类
SubClass	VARCHAR(40)	参看子分类表	子分类
Source	TEXT		数据来源
Publish	VARCHAR(200)		发布机构
PublishDate	Date		发布时间
Size	JSONB		数据量
Footprint	JSONB		
URL	TEXT		发布地址
Images	[]TEXT	URL	数据集图片
Quality	TEXT		数据质量
Zhengce	TEXT		数据政策
Contact	VARCHAR(48)		联系人
ContactTel	VARCHAR(48)		联系人
ContactEMail	VARCHAR(48)		联系人
ContactFax	VARCHAR(48)		联系人

3.2 运行数据

运行数据主要包含两种运行，一个是望远镜的运行数据即观测日志，另一个是数据中心的运行日志。

表 3-5　观测项目元数据

字段	定义	字典	说明
ID	INT8	0-32767	编号
TeleID	[]INT8	0-32767	使用望远镜编号
Name	VARCHAR(100)		中文名称
EnName	VARCHAR(100)		英文名称
Abbr	VARCHAR(20)		缩写
Descrption	TEXT		
Sci	TEXT		科学目标
Start	Date		开始时间
End	Date		结束时间
Footprint	JSONB		
Images	[]TEXT		

3.2.1　观测日志

望远镜的观测日志包括有每次观测的详细信息。

3.2.2　运行日志

数据中心日常运行的日志，包括数据传输、数据访问等。

3.3　服务数据

3.3.1　论文数据

统计记录与数据中心相关的所有数据产出的论文，并从中获取作者、时间、数据、项目等信息。

主要参考 ADS 的论文元数据格式标准，新增字段包括论文下载地址，所属数据集、支持项目等。

表 3-6 论文

字段	定义	字典	说明
ID	INT8	0-32767	编号
Bibcode	VARCHAR(30)		
Title	[]TEXT		
Identifier	[]VARCHAR(100)		
Pubdate	Date		发布时间
Abstract	TEXT		
Readcount	FLOAT4		
Aff	[]TEXT		
Issue	VARCHAR(100)		
Pub	VARCHAR(100)		
Bibstem	VARCHAR(30)		
Volume	VARCHAR(100)		
Keyword	TEXT		
Database	[]VARCHAR(30)		
Author	[]VARCHAR(50)		
Citationcount	INT		
Property	[]VARCHAR(100)		
Page	[]VARCHAR(50)		
Dataset	INT8		数据集
File	TEXT	URL	
Projects	[]INT8		

3.3.2 项目数据

数据中心数据产出论文中提及的各级项目的详细信息，模版可以参考国家自然科学基金的项目信息。

3.3.3 成果数据

从科研成果，由论文、新闻报道中汇总。

3.3.4　新闻报道数据

搜集整理与数据中心相关的所有报道，记录来源，对应数据集、项目等信息。

与数据集、数据中心相关的新闻报道、微信公众号新闻收集，重点与各个数据集关联。

表 3-7　新闻报道

字段	定义	字典	说明
ID	INT8	0-32767	编号
Src	TEXT	URL	来源
Source	TEXT		来源机构
Lang	VARCHAR(4)		语言
Author	VARCHAR(100)		作者
Title	VARCHAR(300)		题目
Time	TIMESTAMP		发布时间
Body	TEXT		
Dataset	INT8		数据集 *ID*
Project	INT8		观测计划

第 4 章　朱雀系统架构

“朱雀”系统是分布式天文数据管理系统的代号。系统的目标是构建一个综合性的天文数据管理系统，以下就系统的基本架构予以简要说明，详细设计参考：《中国天文数据中心分布式存储系统-朱雀-技术手册-v0.1》。

4.1　网络拓扑

整个系统是一个分布式文件系统，组网后的拓扑结构如图4-1，系统有不同的子节点，子节点可以是分布式模式，也可以是简化的小系统部署，甚至只是一个纯粹的客户端，而将管理功能委托给主节点。

主节点是一个分布式架构的系统节点，拥有全部的功能。

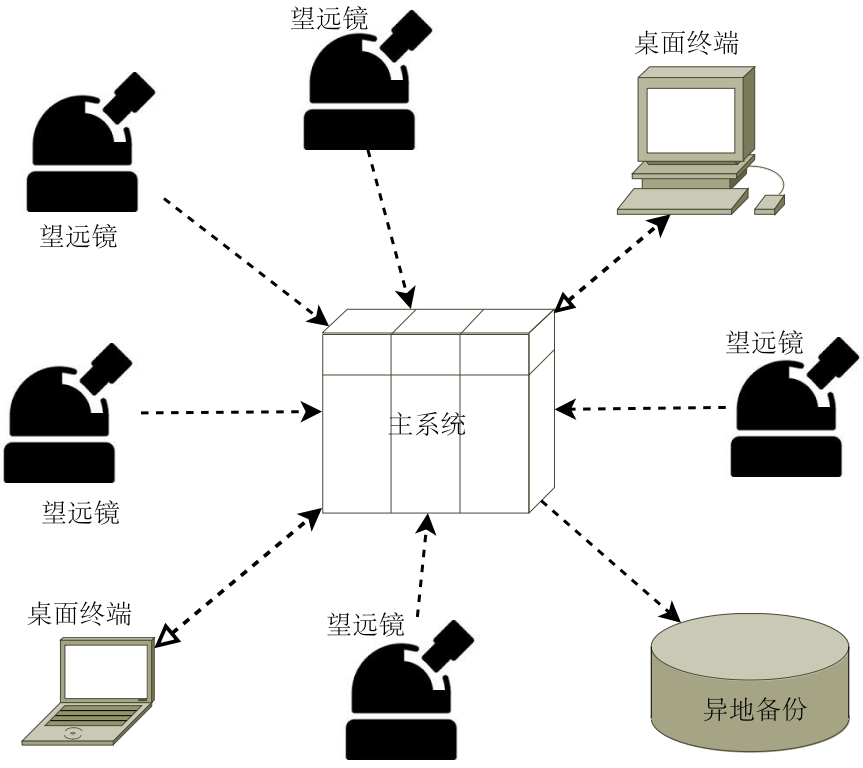

图 4-1　网络拓扑

4.2 系统架构

主系统的整体架构如图4-2，这是一个分布式的架构，主要分为三层：

1. GUI **和客户端**是系统直接针对用户的服务，包括一个用户（管理）界面，针对用户或者客户端的接口，用户可以上传、下载和管理数据。
2. API 是对所有服务的封装，也是对内和对外的一个有效隔离。
3. **分布式系统**是系统的核心组件，包括有

目录、调度服务 提供给用户路由服务，分发用户的上传下载请求至底层的分布式存储服务，并且与中心数据库交互，负责数据集的注册、ID 的命名以及统一管理等。

中心数据库 这是一个高可用的关系型数据库，存储和缓存节点、数据集、数据卷以及数据文件的元信息、这些元信息在数据归档过程中通过逐级汇交等方式汇总在中心数据库中。

运维 包括监控、归档、副本同步等服务。

分布式存储服务 底层存储的核心分布式存储服务，由若干个节点构成，每个节点如图4-8，数据集默认为三副本，并且存储在不同的节点中。

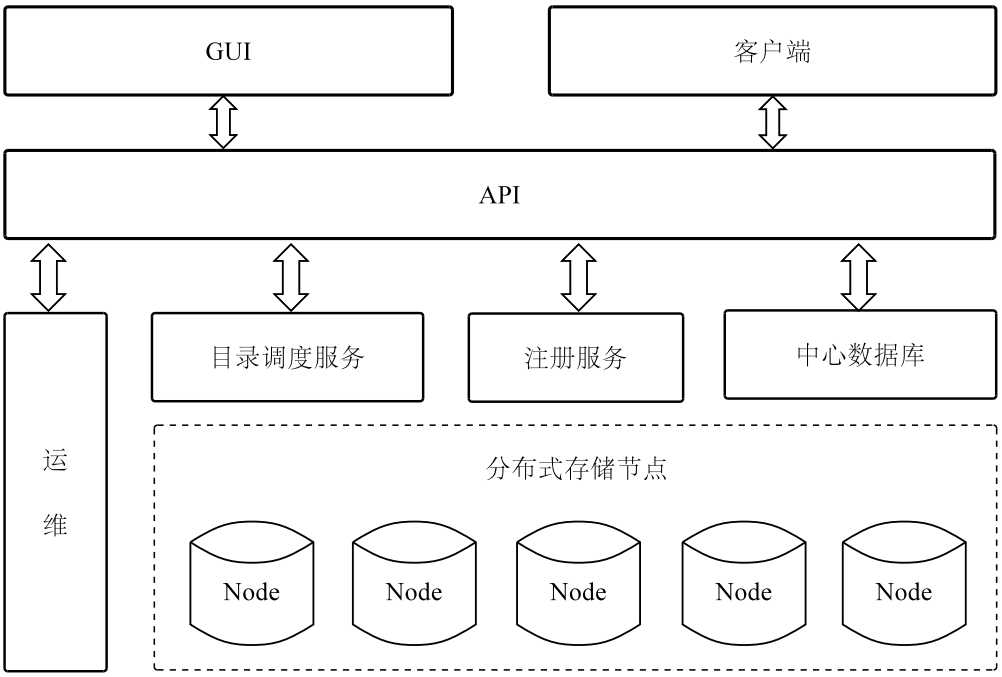

图 4-2 系统架构

4.3　单机模式

根据天文数据管理等需求，将在多个望远镜节点部署服务，因此，可以对系统架构进行简化，如图4-3，形成一个单机版程序。单机版程序即可以在单个服务器部署，也可以在桌面端部署。而桌面端，实际是一个个人管理天文数据的数据管理系统。

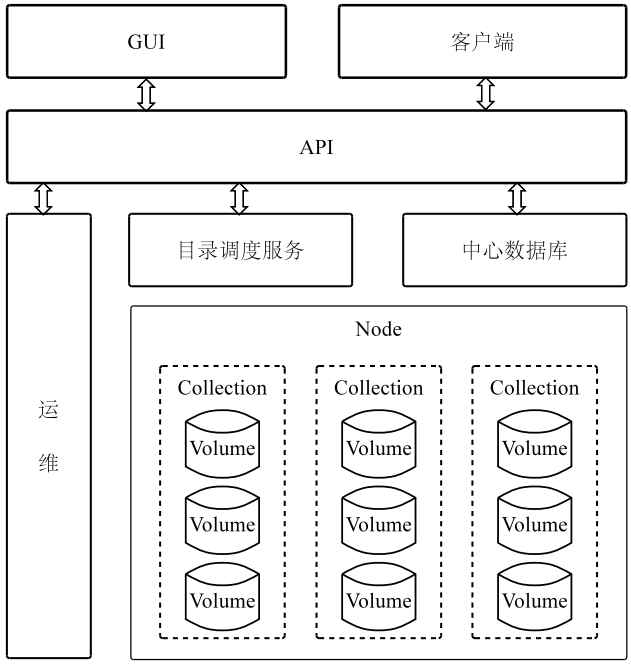

图 4-3　单机模式

系统在设计时就考虑了主系统和小系统（单机模式），这样做的好处是在不同的规模层次使用时，可以选择不同的模式进行部署，并且小系统可理解为分支节点，可部署在望远镜数据中心、个人服务器、桌面电脑上等。并且同时会与中心节点有数据和信息的交互。

4.4　系统功能流程

根据系统总体的架构，如图4-4所示，可以将系统分为对外接口、目录调度、分布式存储、中心数据库、运维几大模块。模块之间通过接口进行解耦，并且可以通过不同的配置组合形成不同层次的系统版本。

4.5　功能模块

根据系统总体架构，如图4-5所示，可以将系统分为五个主要的模块：对外接口、目录调度、分布式存储、中心数据库、运维，这些模块每个又可以分为若干子模块。

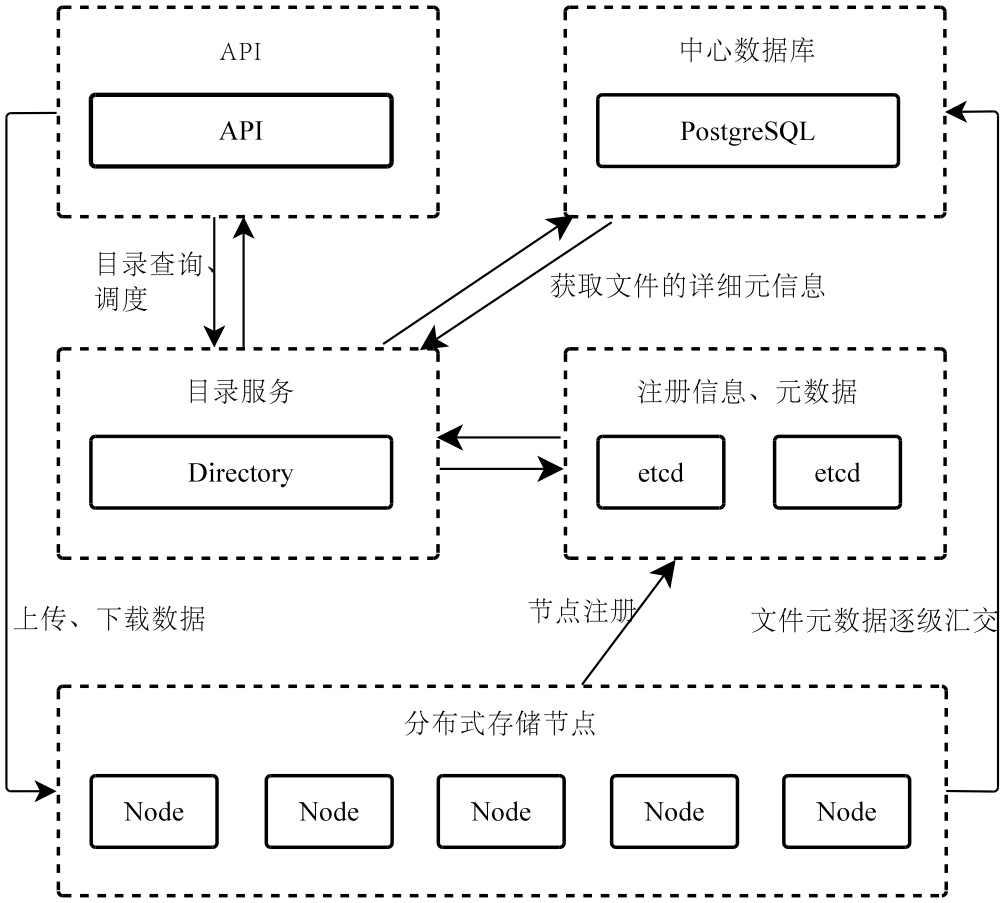

图 4-4 系统功能流程

对外接口

对外接口服务对外屏蔽了分布式系统内部的细节，用户或者前端只需要通过面向资源服务的 RESTful API 就可以使用本系统服务。

对外接口可以分为以下几类服务：

- 注册类：实现数据集的申请、注册；
- 数据操作类：数据的上传、下载、打包；
- 权限类：数据集权限的管理等；
- 统计类：数据、数据集的统计等。

目录调度

目录调度服务是分布式系统的一个核心调度服务，负责将接口的请求进行处理，比如确定节点的位置，文件的信息等。根据请求并且通过接口分发到其他模块进行处理，处理结果再转发给用户。因此这是一个元信息检索定位、存储调度的模块。由于信息都存储在

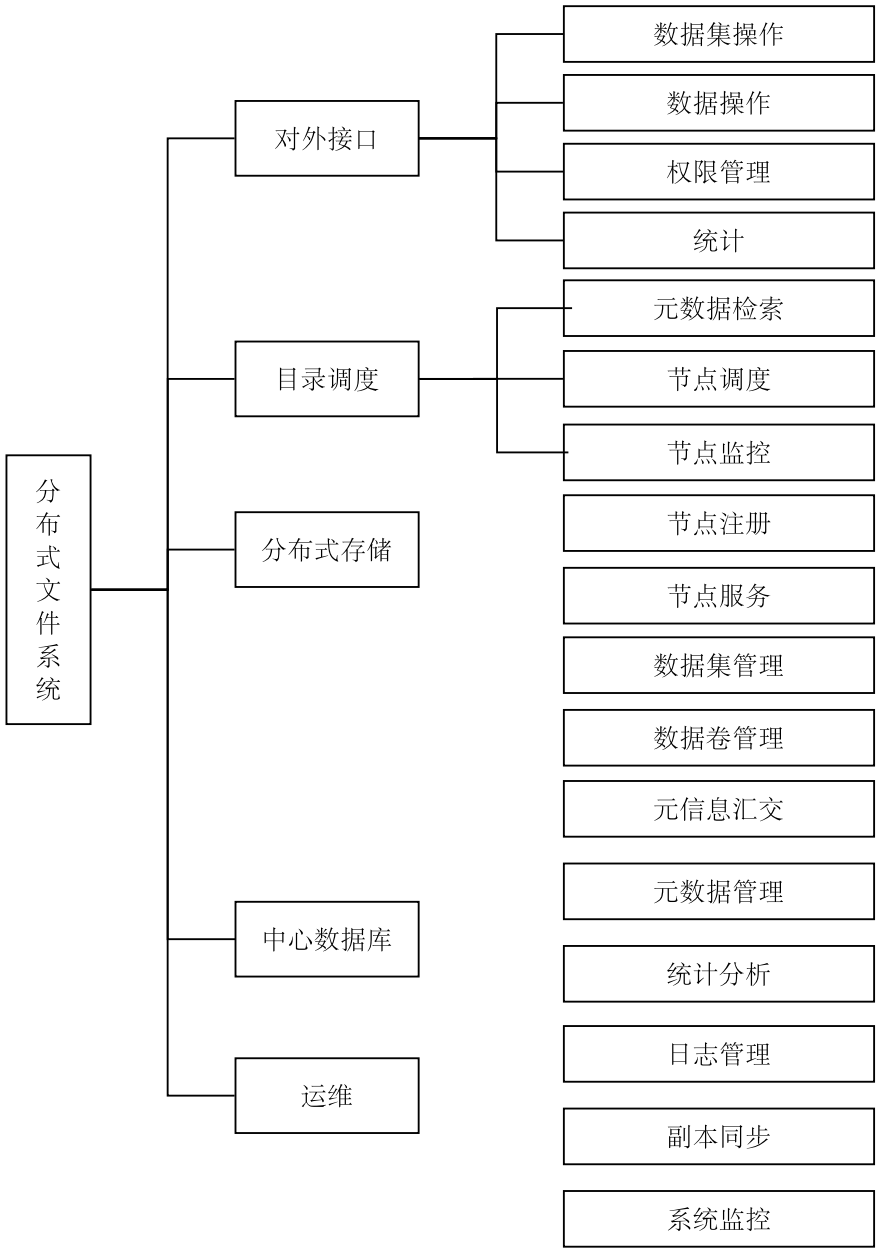

图 4-5　功能模块

中心数据库以及注册服务数据库中，因此目录调度也可以做成分布式多节点的服务，可以根据需要进行扩充。

分布式存储

分布式存储是系统的核心底层服务，由若干存储节点组成。单个节点也可以提供服务。每个节点由多个数据集组成，每个数据集由若干数据卷组成，而每个数据卷则可以存储数千乃至数万的数据文件。

图4-8是一个分布式存储节点的架构图。

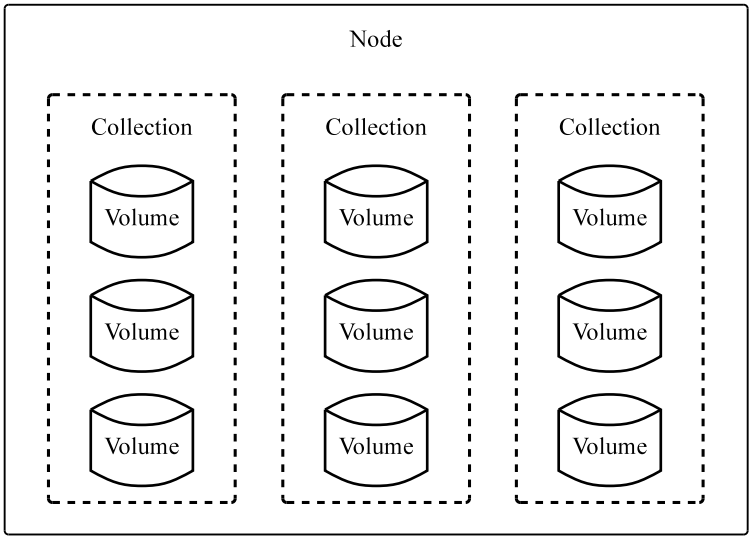

图 4-6　分布式存储节点

中心数据库

中心数据库是一个关系型数据库集群，存储了系统的节点信息、数据集信息、数据卷信息、数据文件信息以及相关的日志信息等。中心数据库通过一个接口实现对数据库的操作，而不是直接操作数据库。

运维

运维模块是一个独立的服务，主要的功能是定期的日志管理汇交，数据集、数据卷等的副本同步等。

4.6　底层存储架构

如图4-7, 底层的分布式存储的结构为一个服务节点（Node）包含多个数据集（Collection），每个数据集由多个数据卷（Volume）组成，每个数据卷由一个超级数据块（SuperBlock）和相关索引文件组成。每个数据超级块由多个基本数据块（Block）组成，数据块是基本的数据单元，存储的是真实的完整数据体，也可以理解为对外表示的一个文件。

4.7　节点数据存储架构

数据节点为一个完整的服务节点。当一个 Node Server 服务启动的时候，将启动对该节点数据管理等服务。

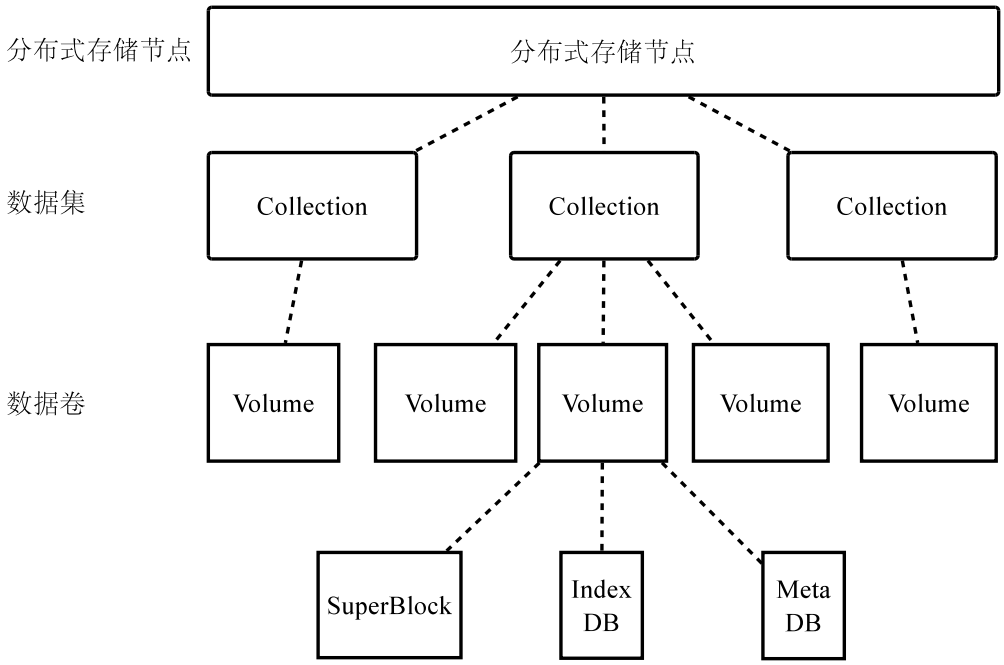

图 4-7　底层存储架构

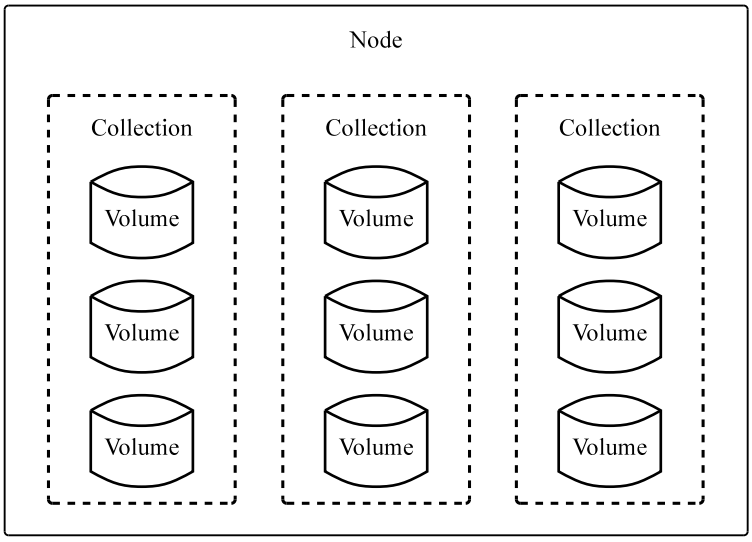

图 4-8　节点

数据节点的目录结构：默认系统数据目录的根节点为`/xingfs`，这个目录也是可以配置的。图4-9就是描述了数据节点的树状目录层次结构。

数据集目录样例：

- `/xingfs/0001E241/` 目录;
- `/xingfs/0001E241/collection.db` 数据集元信息数据库，存储数据集结构和信息。

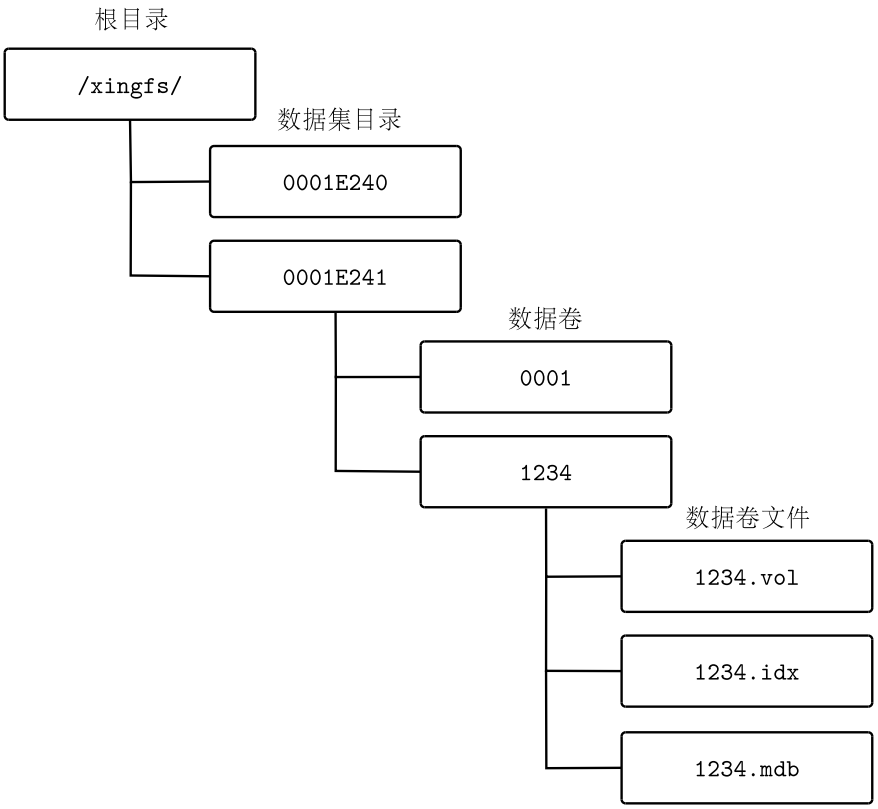

图 4-9　节点数据存储架构

4.8　数据库体系架构

系统数据库体系是一个层级的体系，如图4-10所示，共分为四层，底层是数据卷的数据库，往上是数据集的数据库，再往上是节点数据库，最高层是中心数据库，其中底下的两层可以使用 KV 数据库进行存储，上面两层可以使用关系型数据库进行数据存储。从整个汇交体系上可以看到，每一层都是上一层的一个子集，通过数据的层层汇交，可以实现模块的解耦和数据的高可用。

以上是数据的增量处理的模式，对于数据的删除来说，其流程是将数据块的 flag 置为 0，然后逐级在各层次的数据库中删除记录。对于数据的更新来说，则是先执行删除，再追加新的纪录。

运维系统会定期进行评估，当某个卷的被删除文件量占卷大小的比例小于某个阈值时，进行卷更新操作，卷更新操作是将原卷的有效数据复制到新卷中后，然后删除原卷。

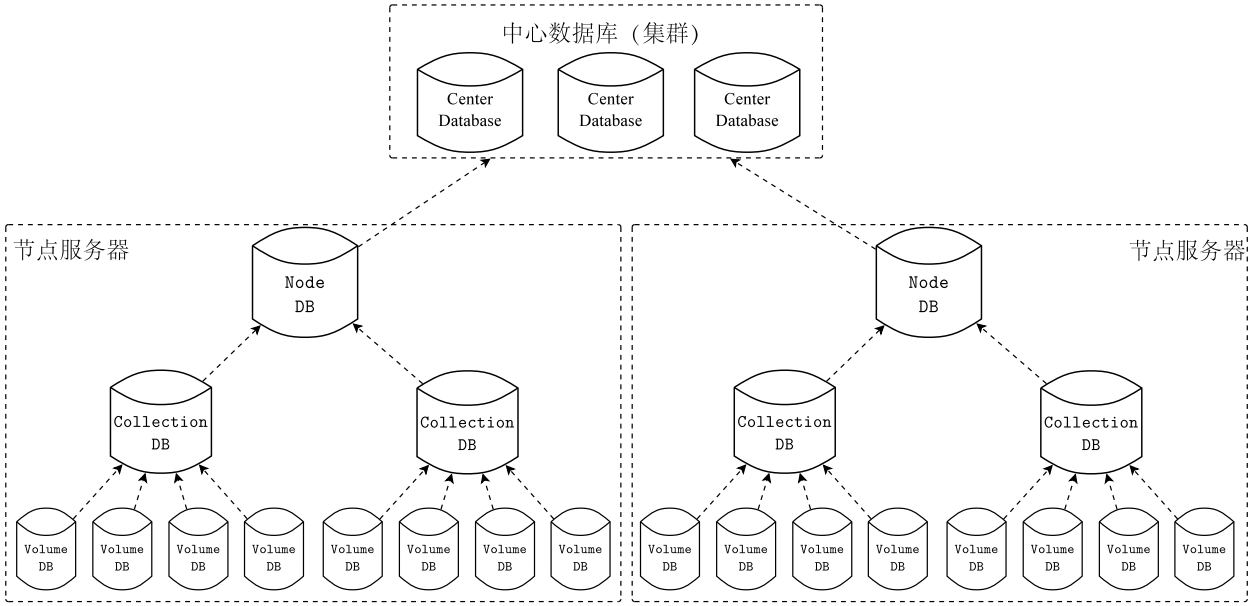

图 4-10　数据库体系架构

第 5 章　结语

路漫漫!