Sparrow 平台方案

华南理工大学 天盛信息科技

1 简介

社会生产与生活均产生大量数据。这些数据通常以电子文件形式进行组织与管理，表现为文本文件（如 doc 文件）、图片文件（如 jpeg 文件）、音频文件和视频文件等。Sparrow 项目作为智能文件管理平台，通过 HTTP REST API 接口形式，为客户管理海量文件（尤其是小文件<15M）提供高效存储、多维检索、智能推荐和安全访问等服务。

2 概念、理念与示例

数据文件的产生及管理与客户业务与组织管理过程密切相关。比如，某企业形成集团、部门和科室的三级层次组织管理结构，并在企业活动中，每个组织通常围绕某个事件或主题产生多个相关数据文件，如某个会议的相关文件，包括会议纪要和会议图片等等。同时，对这些文件的访问还需满足企业业务安全规则，如科室人员未经授权无法查看集团人员产生的文件等。在这个例子中，企业对文件的管理通常不采用扁平结构进行组织，而依据企业组织结构及业务安全规则。

Sparrow 作为文件管理支撑平台，必须对不同客户的具体文件管理行为进行高度抽象，设计并实现通用的管理概念与机制，以便支持客户实际业务需求。Sparrow 的核心概念包括数据组织管理方面的目录（Directory）、文档（Document）和文件（File）、访问控制管理方面的用户（User）、群组（Group）、访问（Access）和许可（Permission）和其他管理概念，包括用户工作区（Workplace）等。

为清晰表述，假设 Sparrow 为一个租户-公司[c1]提供文件管理服务。该组织具有两个部门[d1]和[d2]，及其属下的科室或办公室[o1]、[o2]和[o3]。为进行业务管理，该公司设定公司管理员用户[c1root]，部门用户[d1user]和科室用户[o3user]。在业务安全规则方面，用户[c1root]作为公司管理员用户，对该公司辖下资源拥有完全操作权限，并在业务系统中创建其管理下的其他系统用户；部门用户[d1user]对部门[d1]辖下资源拥有完全操作权限；科室用户[o3user]对科室[o3] 辖下资源拥有完全操作权限；经系统授权，用户[d1user]和[o3user]可对非其辖下资源进行操作。

为满足租户[c1]的需求，Sparrow 构建如下图 1 所示概念结构。

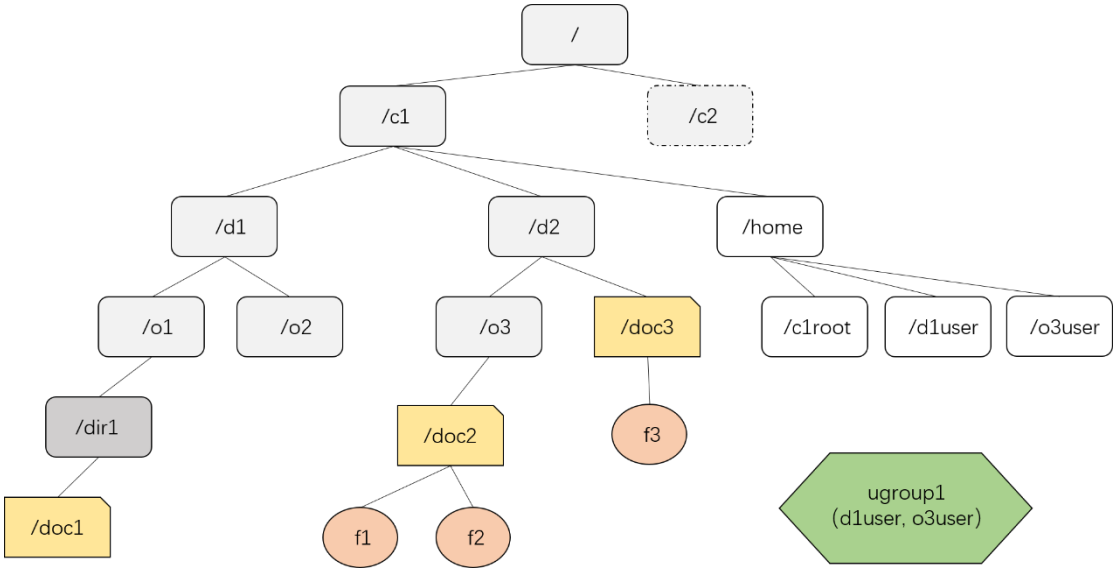

图 1 概念结构示例

- **目录（General Dir，GDir），用于构建逻辑“文件”系统，可映射为客户树形层次组织管理结构等。**

在计算机或相关设备中，一个目录就是一个装有文件系统的虚拟“容器”。在它里面保存着一组文件和其它一些目录（又称子目录）。这样，这些目录（及其子目录）就构成了层次（hierarchy），或树形结构。

在 Sparrow 中，前述定义中的“文件”对应其管理概念文档，而不是具体数据文件。

在 Sparrow 当前版本中，仅支持一个目录树，即所有子目录从根目录“/”派生。

示例图 1 中，灰色圆角四边形为目录对象，如[o1]，其路径为[/c1/d1/o1]。

示例图 1 中，为租户[c1]创建了如下目录结构：

➢ 机构型目录（Organization Dir, ODir）：

1. [/c1]：为该租户[c1]的根目录；
2. [/c1/d1]和[/c1/d2]：为相应部门根目录；
3. [/c1/d1/o1]、[/c1/d1/o2]和[/c1/d1/o3]：为相应科室的根目录。

➢ 用户 home 目录（Home Dir, HDir）：

1. [/c1/home/c1root]：公司管理用户[c1root]的默认 home 目录；
2. [/c1/home/d1user]：部门[d1]用户[d1user]的默认 home 目录；
3. [/c1/home/o3user]：科室[o3]用户[o3user]的默认 home 目录。

➢ 业务型目录（Business Dir, BDir）：

1. [/c1/d1/o1/dir1]：随业务开展，用户在目录[o1]中创建了子目录[dir1]。

- **文档，用于构建逻辑文件管理集合，映射客户业务过程中的事件或主题等。**

在 Sparrow 中，文档不是具体的数据文件，而是管理文件的容器。文档是一种特殊的目录（Document Dir，DocDir），其必须从属唯一的父目录，其子节点只能是下文所述的文件。

示例图 1 中，橙色缺右上角四边形为文档对象，如[doc2]，其路径为[/c1/d2/o3/doc2]。

- **文件，指具体数据文件，如一张图片等。**

在 Sparrow 中，文件指具体数据文件，比如一张图片等，其必须从属于一个文档对象。

示例图 1 中，红色圆形为文件对象，如[f3]，其路径为[/c1/d2/doc3/f3]。

- 工作区（Workplace，WP）：业务工作区（Business WP， BWP）和个人工作区（User WP，UWP）；

为安全访问目录、文档和文件等核心资源（Resource），Sparrow 采用面向资源的操作许可访问控制策略（Resource-oriented and Permission based），同时支持两种具体机制：1）扁平化权限管理策略（类 Linux OS 文件权限管理），和 2）层次权限继承的管理策略。其访问控制策略涉及如下概念：

- **用户，指系统使用者，是目录、文档和文件的操作者。**

  Sparrow 初始化后，默认存在一个全系统域 root 用户，拥有整个系统资源的全部权限，并在业务过程中，创建其他系统用户。

  示例图 1 中，为租户[c1]共创建了如下用户：

  1. [c1root]：租户[c1]的默认管理员账户；
  2. [d1user]：部门[d1]用户[d1user]；
  3. [o3user]：科室[o3]用户[o3user]。

- **群组，指一组用户的集合，便于授权管理。**

  Sparrow 通过将一组用户构成一个群组，便于授予该组用户对一组资源相同的访问许可。

  示例图 1 中，群组[ugroup1]包含用户[d1user]和[o3user]，群组中的用户继承群组的权限。

- **访问，对资源的操作，包括读取、创建、删除和执行（待设计）。**

  （当前版本：采用 Linux 的读、写和执行）

- **许可，用户或群组拥有对资源具体操作的准许，授予许可的过程称为授权。**

  示例图 1 中，用户[o3user]对文档[doc2]拥有读取操作许可，即读权限；群组[ugroup1]对文档[doc3]拥有读取操作许可。

3 系统设计

在设计上，Sparrow 必须实现数据组织类资源（目录、文档和文件）的高效管理，安全访问的灵活控制，并支持多租户、分布式部署。（当前版本：不支持多租户和分布式部署）

3.1 总体结构

Sparrow 采用模块化与层次结构，如下图 2 所示。

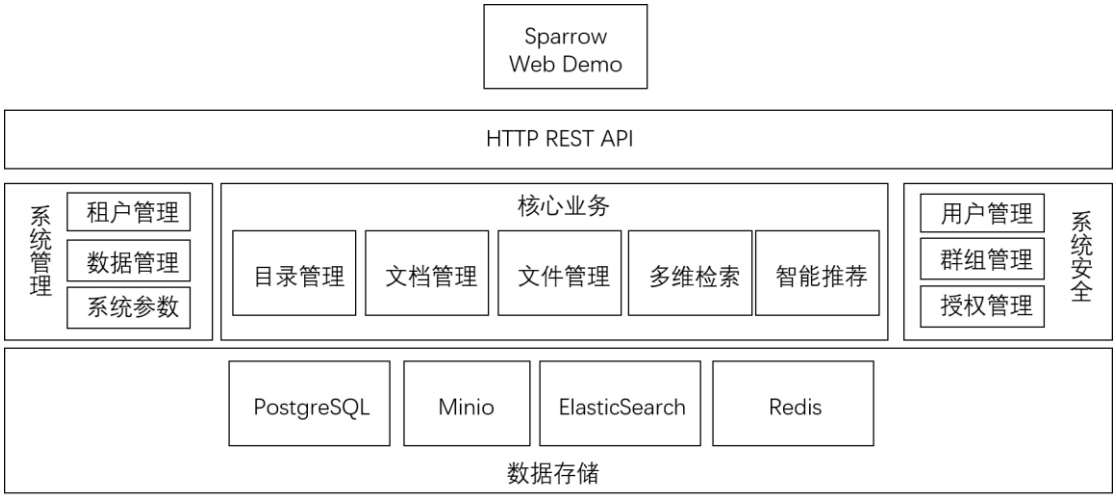

图 2 系统结构

Sparrow 主要包含系统管理、系统安全、数据存储和 HTTP REST AIP 服务模块。

系统管理，包括 1）系统参数配置管理；2）数据管理，支持用户定义文档和文件的元数据模型、多维检索及智能推荐模型；3）租户管理，支持多租户。

系统安全，包含 1）用户管理，2）群组管理和 3）授权管理。

核心业务，作为文件存储、检索与推荐的核心业务，包含 1）目录管理，2）文档管理，3）多维检索和 4）智能推荐。

数据存储，业务数据的持久化，综合采用了关系型数据库 PostgreSQL，面向对象的存储 Minio，全文检索引擎 ElasticSearch 和 Key-Value 缓存数据库 Redis。

其中，系统管理、系统安全和核心业务模块，通过 HTTP REST API 接口提供服务。Sparrow Web Demo 是为演示系统核心概念和功能的示例性应用。

3.1.2 数据模型

数据模型指系统业务对象及业务过程临时业务实体的属性表达。这里仅描述业务对象，临时业务对象通常产生于具体服务过程，将在相应服务环节进行介绍。

系统状态可简单视为其全部数据模型的当前值状态，业务操作的结果可能产生系统状态转变，并最终实现系统状态的持久化。数据模型是一个抽象概念，其描述了具体属性的类型等约束，与具体持久化机制无关；但不同的持久化方案，可能产生不同的效率和可维护性。因此，通常在设计数据模型应综合考虑目标持久化机制。

如下数据模型，需采用表格形式说明其类型等规范，并给出与示例图 1 中相关的实例。

目录：一种树形结构的节点（PostgreSQL，说明其采用 PostgreSQL 持久化）（后续支持目录类型： oDir， hDir 和 bDir 等）。

文档：在 Sparrow 中存在两种数据模型；作为目录结构中的对象（PostgreSQL），和支持全文检索的对象（ElasticSearch）；

文件：注意，虽然示例图 1 中，文件看起来作为目录树的叶子节点，但是文件不属于目录结构。文件是全文检索对象（ElasticSearch，说明其采用 ElasticSearch 持久化），其与文档的关系在文档的 ElasticSearch 对象中表达。

注意，文档和文件作为 ElasticSearch 全文检索对象，除了系列号、创建者、创建时间、修改时间、标题、类目、标签、关键字和简介属性外，其他属性应有租户根据业务设定。

租户：（PostgreSQL）

用户：（PostgreSQL）

群组：（PostgreSQL）

许可：（PostgreSQL）

3.2 核心服务

海量数据文件高效管理是本质需求。为此，Sparrow 提供高效存储、多维检索、智能推荐和安全访问服务。

3.2.1 高效存储

为实现文件高效存储，Sparrow 采用目录及文档构建逻辑树形层次结构，一方面，层次结构便于合理组织与管理数据，比如备份某个目录下的文件，比备份整个系统所有文件，应该涉及更少的文件，另一个方面，可通过初始化机构类型目录，映射客户组织结构，然后组织成员可在其所属机构目录下进行目录、文档和文件管理操作。

文件的高效存储，由目录、文档和文件的 CURD 操作实现。

Sparrow 基于 PostgreSQL 实现目录（郑铠锋），Minio 实现文件存储（陈晓滨、梁宏达），ElasticSearch 存储文档和文件的数据模型数据（元数据，Meta Data）（陈绿佳）以提供多维检索功能。

在上传文件后，Sparrow 支持文件缩略图生成；如果是视频文件，支持截取若干帧图片，采用 base64 编码存储（不建 ElasticSearch 反向索引），后续支持图片检索。对文本类型文件，如 doc，转换为 pdf 后，再生成缩略图，采用 base64 编码存储（不建 ElasticSearch 反向索引）。

3.2.2 多维检索

从检索对象角度，Sparrow 支持检索目录、文档和文件。

目录检索：假设某目录包含 50 多个子目录或文档，系统支持在目录下检索名称包含检索关键字的目录（或文档）；可采用客户端，如 Web 浏览器，进行字符串模糊检索实现；

文档检索：构建文档数据模型的 ElasticSearch 索引，提供多维结构化检索与全文检索。

文件检索：构建文件数据模型的 ElasticSearch 索引，提供多维结构化检索与全文检索。

为支持多维检索，Sparrow 构建了 4 个 ElasticSearch 索引库（Index），图示如下：

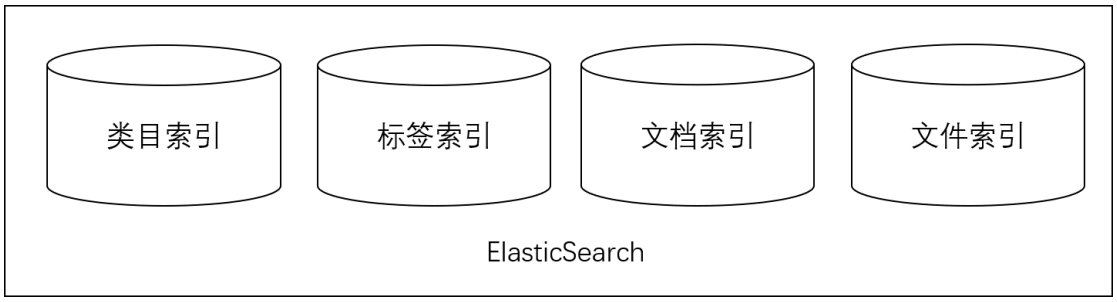

图 3 ElasticSearch 索引

文档检索与文件检索，待分析陈绿佳的设计。

3.2.3 智能推荐

待设计与整合!! 基于内容的统计，和协同推荐！待设计!

3.2.4 安全访问

（当前版本：不对文件对象进行授权，而是对其父节点文档进行授权，即用户对文档的操作权限自动传递给文档内所有文件）

在 Linux 文件权限管理模型中，资源的创建者成为该资源的拥有者（Owner），默认拥有对该资源全部访问权限；系统其他用户作为该资源的其他用户（Others），默认不拥有对该资源的任何权限。Root 用户和资源拥有者，可对某用户授予对该资源的访问操作许可，也可先创建一个普通权限用户群组（Regular Group，RG），再对该群组授予对该资源的访问操作许可（RG Permission，RGP）。

Sparrow 支持上述访问控制机制。该机制粒度细，灵活，适配场景完全，但是授权复杂，成本高。因此，Sparrow 还同时支持另外一种层次结构授权与访问控制机制，实现与客户层次组织权限管理结构相结合的访问控制，提高授权效率，降低成本。该层次结构访问控制只针对目录资源进行管理，并由 Root 用户或其他授权进行统一实施。Root 用户可先创建一个层次权限用户群组（Hierarchical Group，HG），再对该群组授予对某目录资源的访问操作许可（HG Permission，HGP）；HG 群组对该目录及其任意深度子目录均具有被授予的访问权限。

（当前版本：部分支持，需要重新设计，参考 ThingsBoards 进行设计与实现）

（注：需要结构图进行详细说明！）

（注：细化设计，理顺程序判断流程！）