



金融行业存储精益化管理，降低存储成本

王凡

资深解决方案顾问

CONTENTS

- PART 1 存储管理现状与挑战
- PART 2 存储精益化管理解决方案
- PART 3 工作实施内容及优化成效

PART 01

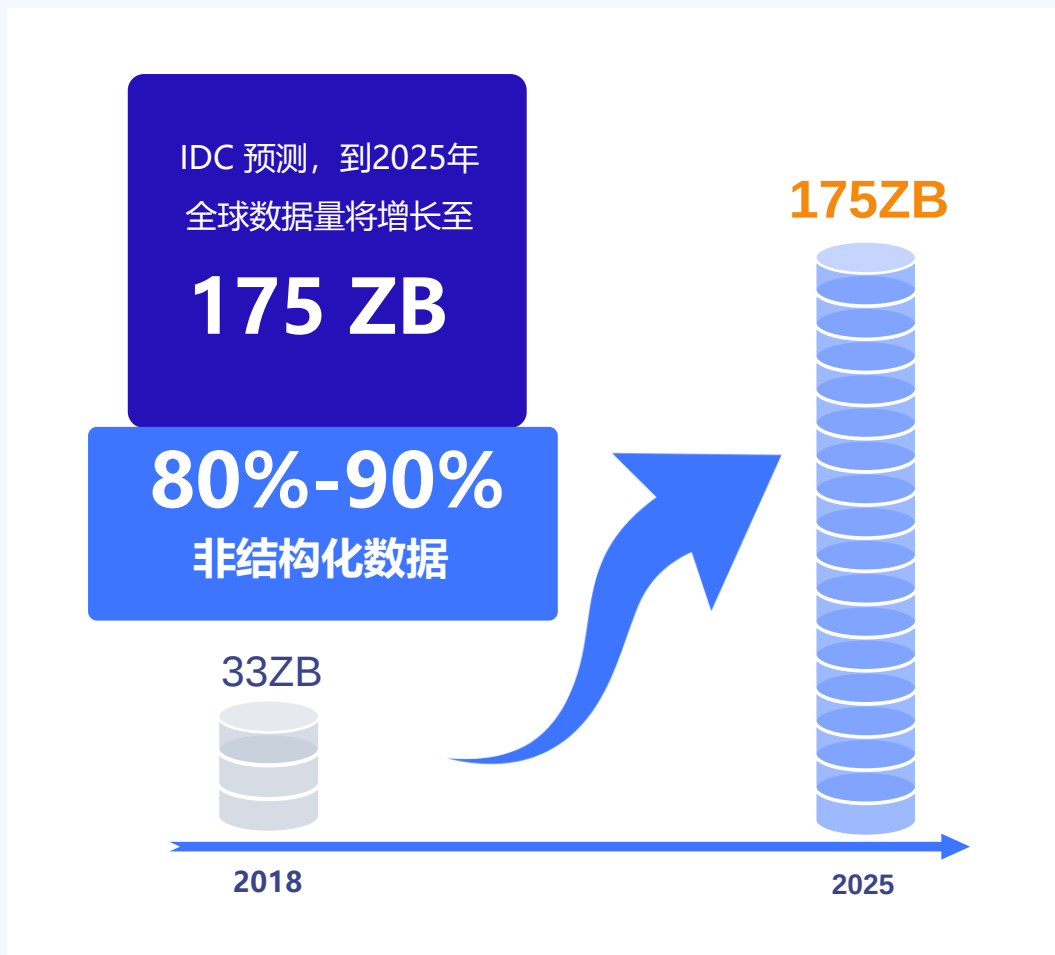
存储管理现状与挑战

现状：近年银行存储采购情况分析

- 不同规模的银行根据其业务需求和预算，在存储设备上的投资金额相对比较大，从小型的几百万元到大型的几千万元不等
- 金融行业通常是中低端存储和高端存储搭配使用，集中存储和分布式存储的混合使用，最终的结果是通过架构的设计或管理要求，既要保证系统的高可用性和性能要求，又要用最低的成本采购到最适合的存储设备

银行	项目名称	金额（元）
人行征信中心	2024年存储采购项目	2211万
兴业银行	2022年存储云化设备扩容项目	3444万
人行征信中心	2024年存储采购项目	2211万
浦发银行	2024年信用卡中心存储资源池采购	1036万
中国银行	2024年总行IT基础硬件资源池分布式存储	966万
北京银行	2024年蓝光存储采购项目	130万
	2024年分布式存储采购	142万
	2024年信用卡机房统一存储平台	181万
邢台银行	2024年存储设备采购	365.8万
九江银行	2024年存储设备采购项目	240万
绍兴银行	2024年分布式存储采购项目	196万
哈尔滨银行	2024年NAS存储采购	192万
广东农商行	2024年存储替换纸全闪存储采购项目	135万
注：数据来源于采招网公开信息		

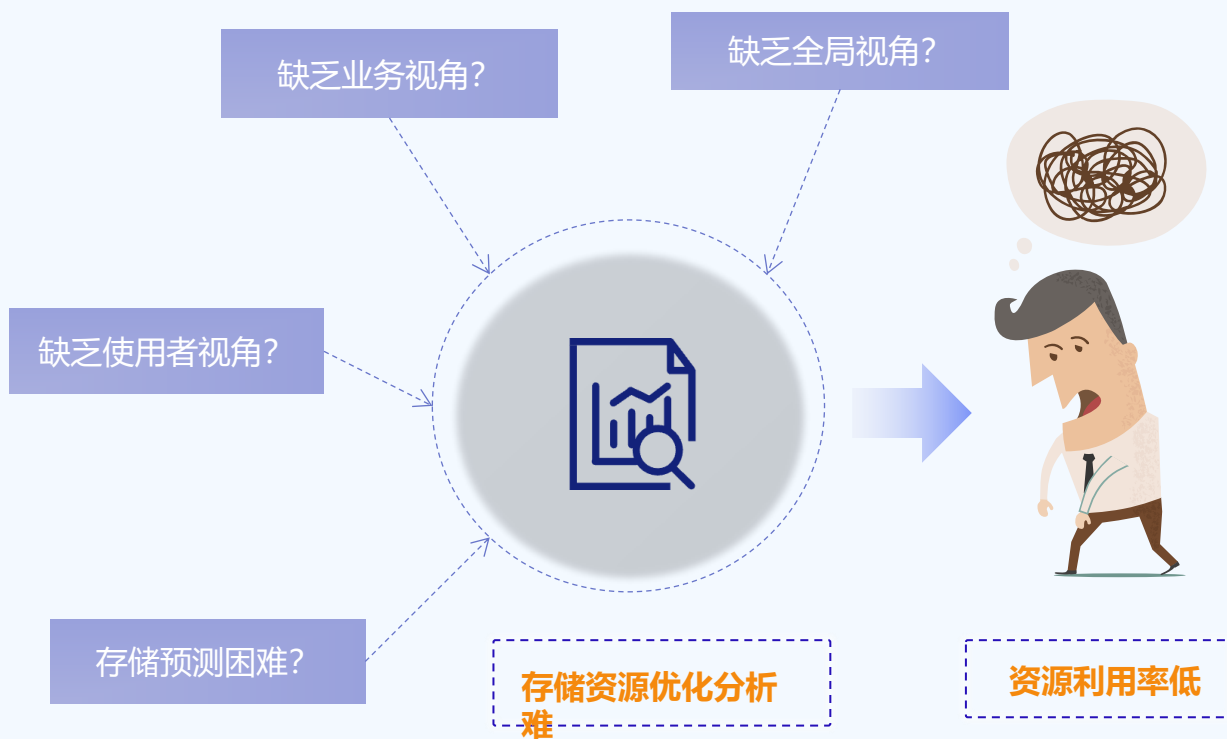
挑战一：海量的数据增长，存储年年买，成本居高不下



- 随着数据量越来越大，对存储的需求越来越大，存储成本逐年升高
- 非结构化数据量占比逐步增加
- 存储数据未分类，大量“冷数据”与频繁修改、点查的数据放在同一环境中进行同等处理

- **数据无序堆积**：随着各种非结构化数据（文本、图片、视频、日志等）在存储平台间的不断膨胀，数据的存储成本居高不下。
- **存储资源浪费**：大量重复、过期文件占用宝贵的存储空间，费用高昂。
- **成本控制困难**：缺乏有效的数据管理和优化策略，难以分辨哪些数据必须长期保留，哪些数据是冷数据可以丢弃，哪些数据是热数据需要存储在高性能但高成本的存储介质上，哪些冷数据可以迁移到低成本的存储介质上。

挑战二：存储资源利用率普遍较低，缺少多视角的存储资源优化分析



- **缺少全局视角：**企业往往需要跨多个平台甚至地理位置整合存储资源。缺乏全局视角使得企业难以对存储资源进行全面的监控和优化，导致资源利用率低。
- **缺少业务视角：**多样化的业务需求导致各企业需要在不同业务单元之间平衡存储资源，以满足各自的性能和容量需求。但是，大部分企业缺少业务视角的资源使用分析，难以将存储资源的分配与业务需求和优先级对齐。
- **缺乏使用者视角：**在企业为员工提供虚拟化桌面的背景下，员工是存储资源的直接使用者，他们的个人存储空间管理对整体资源利用效率有着直接影响。然而，许多企业在存储资源优化分析时往往缺乏对使用者需求的考虑，导致存储资源的浪费和成本不必要的增加。
- **存储需求预测困难：**业务数据量迅速增加，导致存储需求不断上升，但是大部分组织的存储需求预测缺乏前瞻性，可能导致供应不足。

挑战三：存储数据资源难以梳理，且管理不足



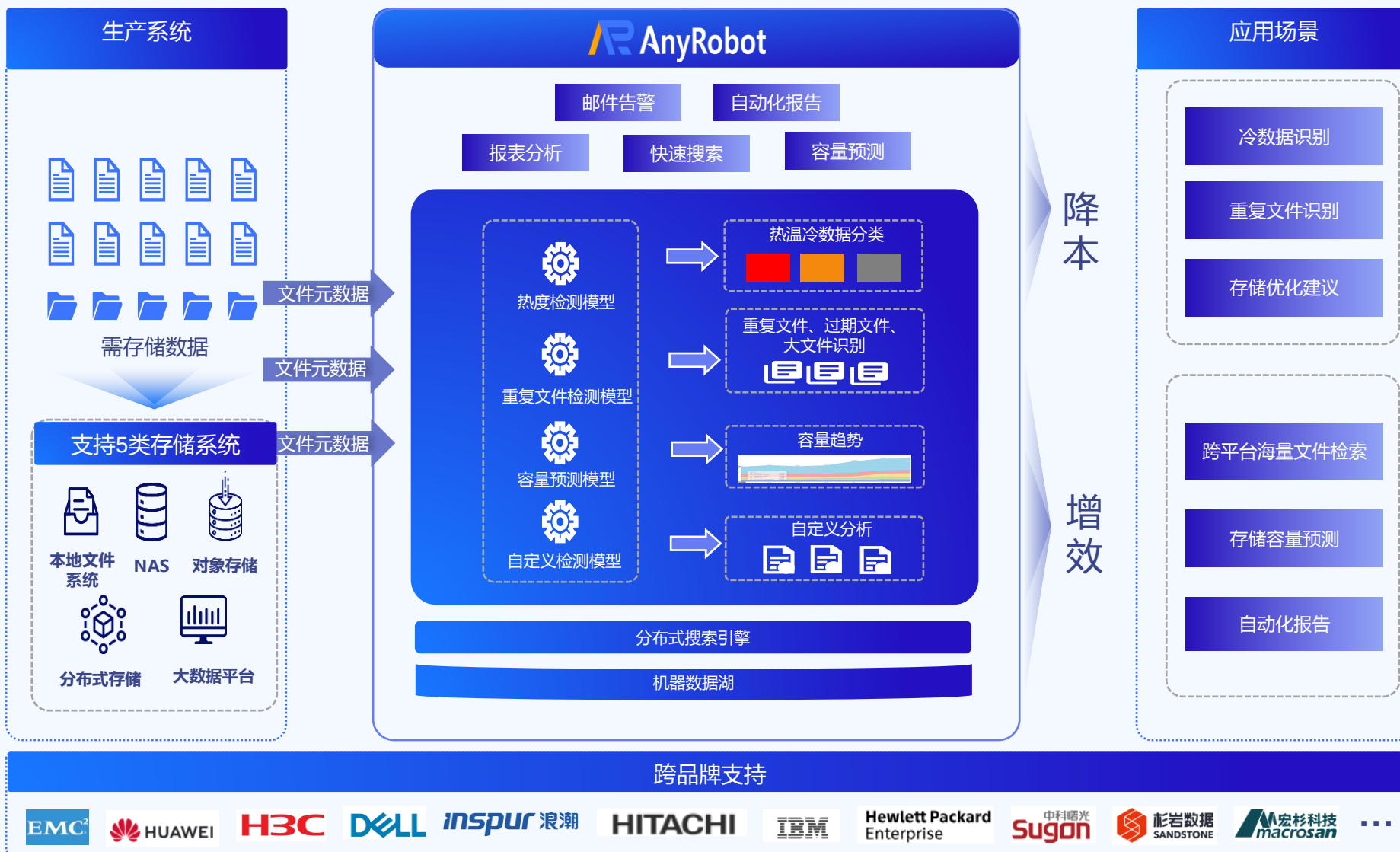
- **多样化的数据类型:** 客户拥有包括文本文档、图片、视频、音频和日志文件等在内的多种类型的非结构化数据。
- **跨多个业务部门的运作:** 企业的各个部门通常都有独立的系统来管理其业务流程，并生成各自的数据，这些数据往往被封闭在各自的系统中，缺乏共享机制和统一管理与分析的能力。
- **不同用户需求的差异:** 企业的不同用户对于数据的使用需求各不相同。一些用户需要全局的可视化数据报表，另一些则需要详细的实时数据报表，需求的多样化导致数据在整合和分析过程中变得复杂。
- **异构的存储环境:** 数据分散存储在不同的存储系统，如网络附加存储(NAS)、存储区域网络(SAN)、对象存储等。这些环境各自使用不同的协议、标准和访问控制机制，加剧了数据的分散，导致难以实现跨平台的数据汇总和分析。
- **缺乏统一的数据视图:** 企业无法全盘掌握其所有数据资源。这意味着某些数据可能被忽略或未被充分利用。

PART 02

存储精益化管理解决方案

存储精益化管理解决方案架构

- **跨平台跨品牌统一管理：**
支持5种主流存储系统，跨品牌支持异构存储全景管理，打破数据孤岛
- **冷数据分析识别：**
内置多种模型与算法
- **数据资源梳理：** 内置多种数据类型识别、分类、评估模型，优化数据分层管理
- **重复、过期文件识别：**
识别、合并或删除重复文件
- **存储优化建议：** 提供多种自动化报表，针对存储优化建议
- **预测性存储规划**
趋势分析、资源预测
- **PB级文件分析**
分钟级处理



通过存储精益化管理，支撑数据分级管理



文件数据自动分类

内置多种分类模型，包括文件热度、文件类型、业务属性的自动分类；支持自定义分类规则，可根据自定义标签、正则表达式等对文件进行自动分类。

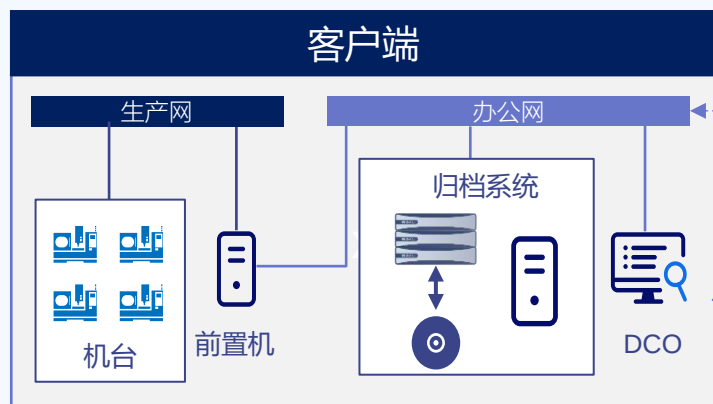
优化数据管理策略、提高数据可靠性

对文件进行热、温、冷分类后，管理员可以根据数据的分类情况制定相应的存储策略、备份策略、恢复策略等，保证数据的可靠性和安全性。

数据分类存储、降低成本

通过将温数据、冷数据存储较低成本存储介质上，实现数据分类存储，从而降低成本、提高系统性能。

自动化数据转储和删除，减少操作风险、降低总体存储成本

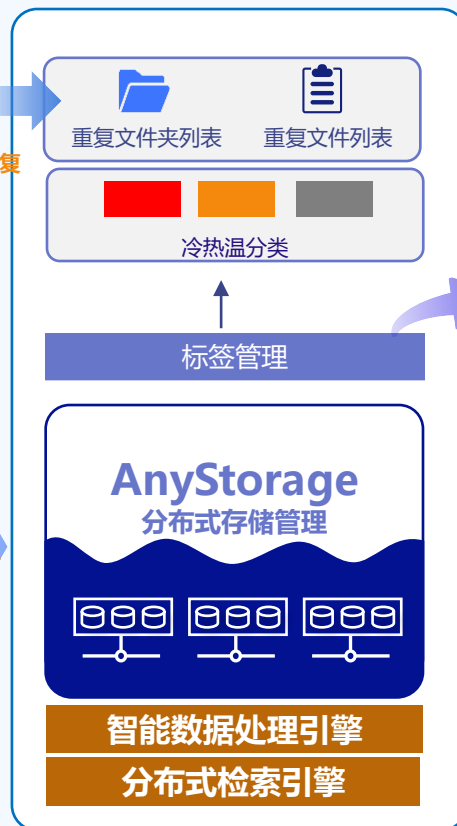


NFS挂载 | CIFS挂载 | 本地Agent | REST API
文件元数据读取

AnyRobot



待转数据处理



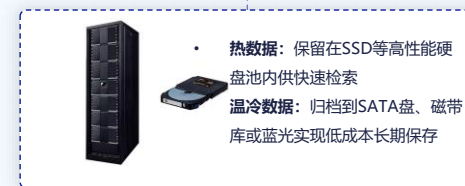
重复数据
可删除/可恢复

压缩和重复数据删除

支持压缩或删除重复文件，降低冗余数据所需的存储量，并降低成本。



自动化数据转储



分层归档



自动化热温冷分级存储，降低存储成本

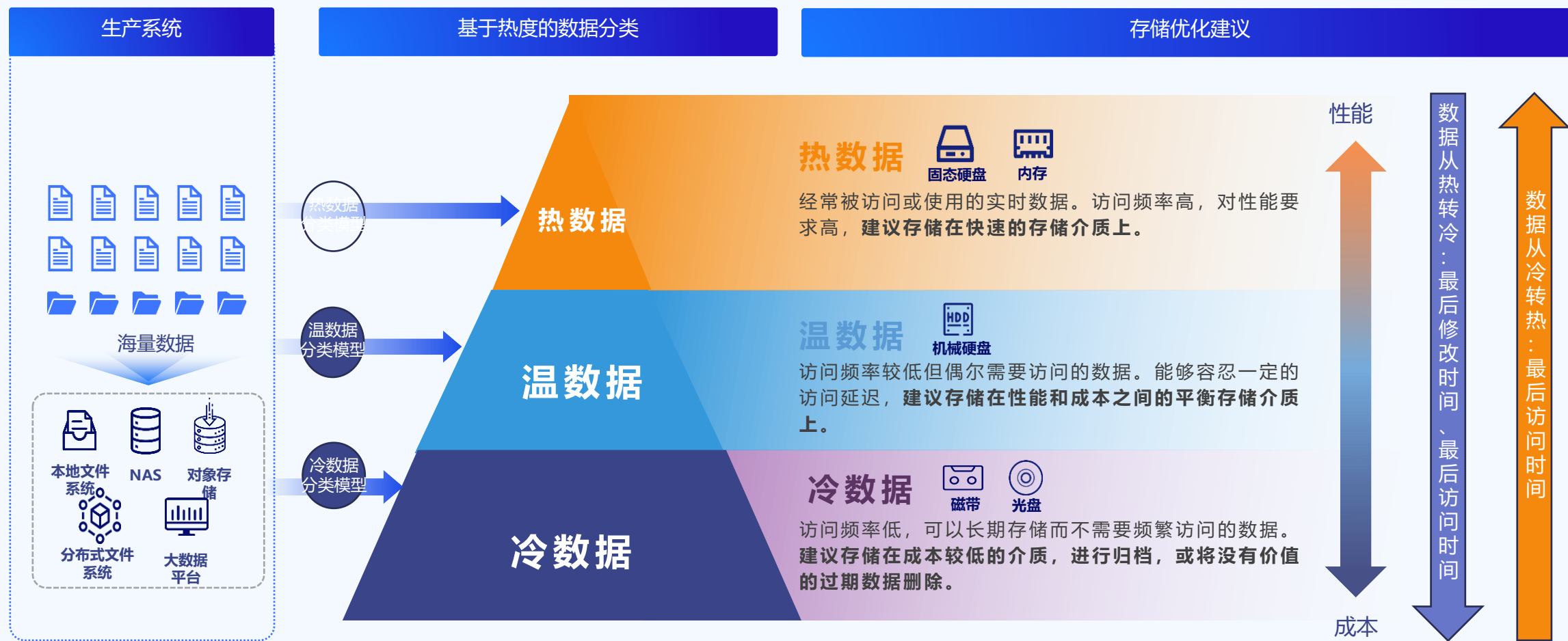
支持动态调整冷热温数据的存储策略，将热数据迁移在快速存储介质上，温冷数据迁移到成本较低的存储上，降低降低存储成本。

PART 2.1

降低存储成本

- 冷数据识别
- 重复文件识别
- 存储优化建议

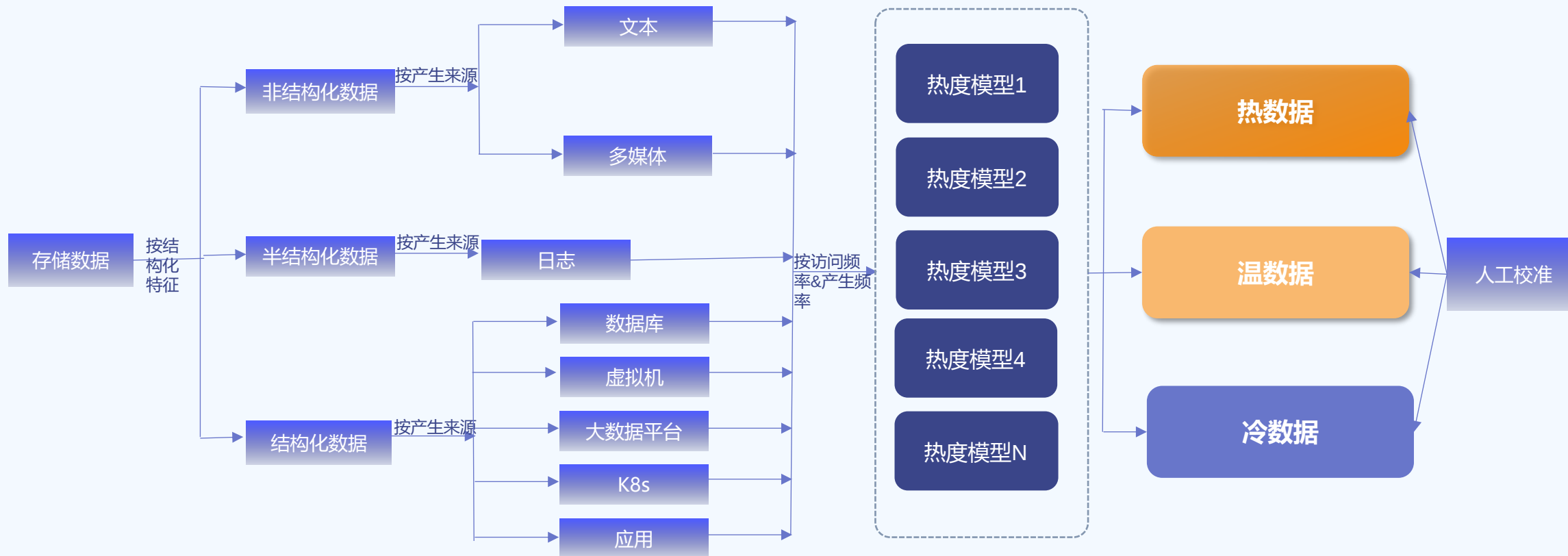
冷数据识别，提供数据迁移及转储依据



- 支持创建时间、最后访问时间、最后修改时间的冷数据识别
- 支持识别模型的自定义

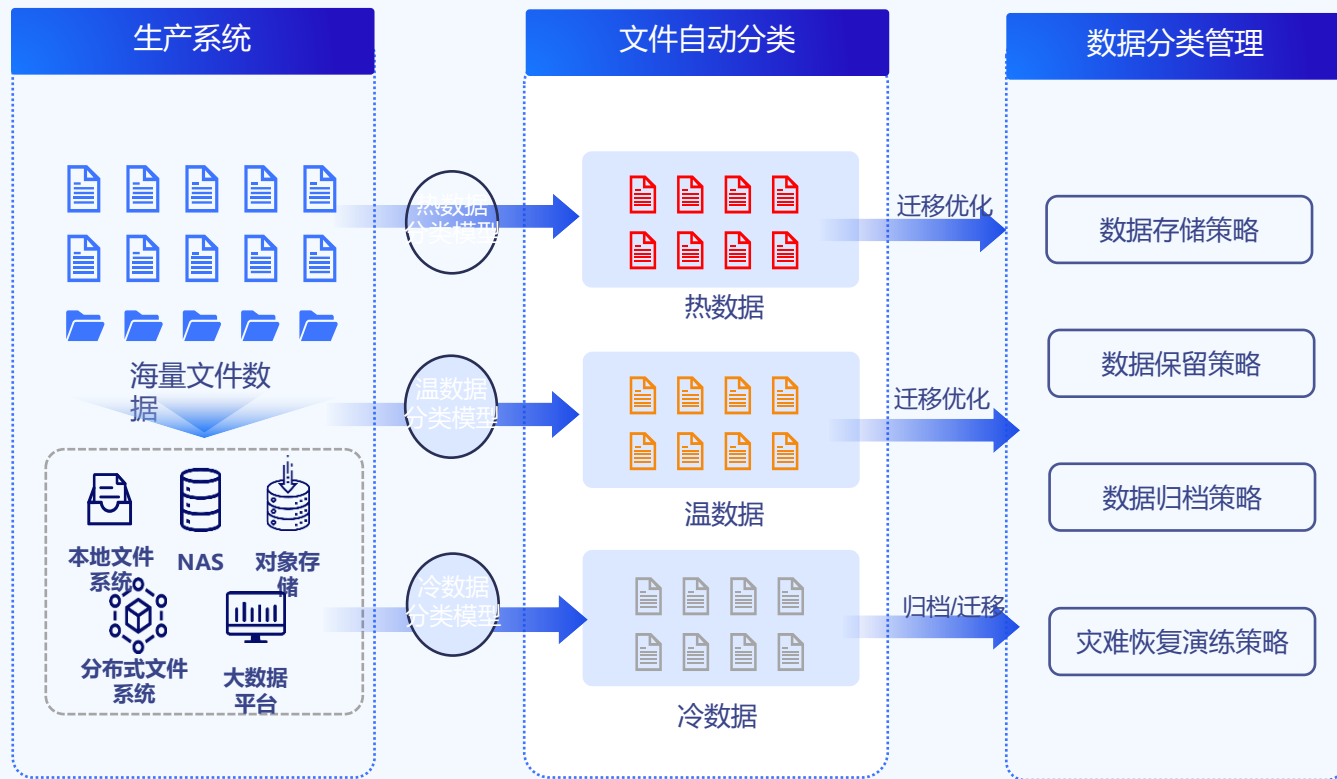
冷数据评估模型与规则

按照数据的创建时间、最后访问时间、最后修改时间，将数据分为冷、热、温三个类别。



注：依据《GB/T 38667-2020 信息技术 大数据数据分类指南》分类维度及线性分类法

文件热温冷分类，优化数据分层管理



场景描述

随着数据量的增长，如何在保证数据可靠性和可用性的前提下，降低存储成本，是一个重要的考虑因素。不同的业务对数据的访问频率和重要性有不同的要求。在线业务系统需要实时访问和处理热数据，离线分析系统则需要处理大量的历史数据和冷数据。需要通过对数据进行热、温、冷分类，根据数据的访问频率和价值，选择不同性能和成本的存储介质和存储方案，实现存储成本的优化、满足不同业务对存储系统的需求。

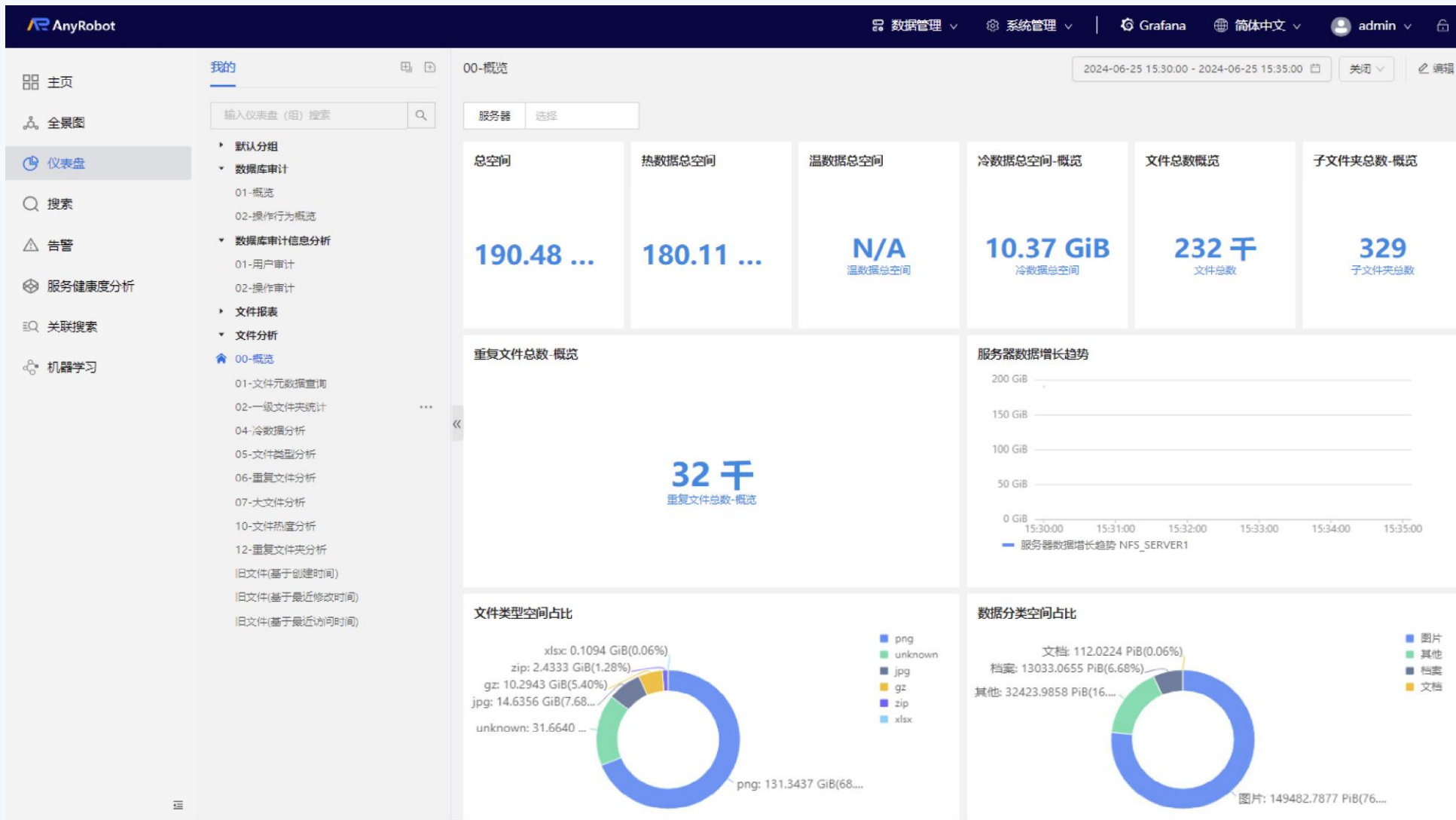
业务价值

- **冷数据迁移、降低成本：**通过将温数据、冷数据存储较低成本存储介质上，实现数据分类存储，从而降低成本、提高系统性能。
- **优化数据管理策略、满足不同业务需求：**对文件进行热、温、冷分类后，管理员可以根据数据的分类情况制定相应的存储策略、备份策略、恢复策略等，保证数据的可靠性和安全性。

存储数据的全局视角管理

存储全局概览

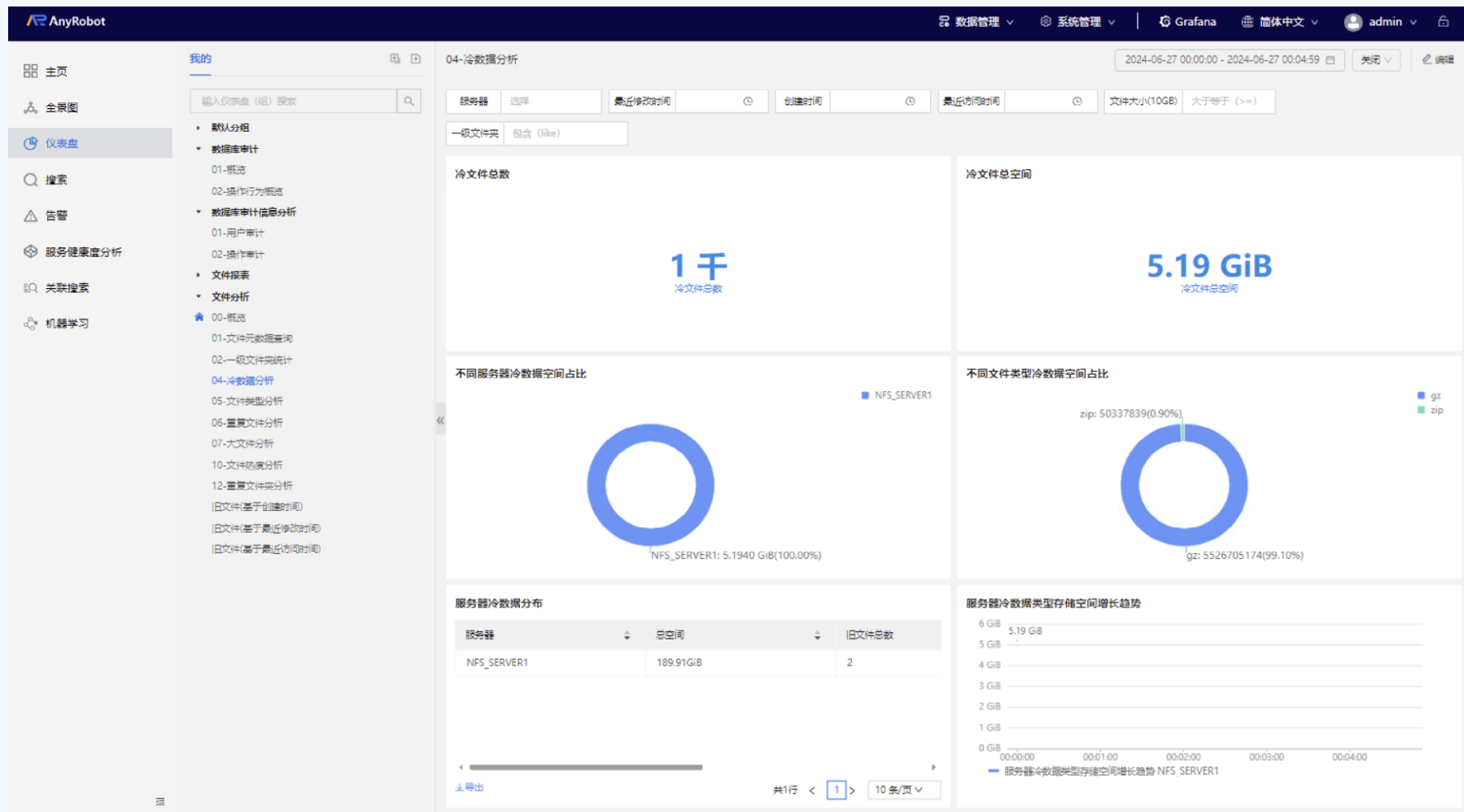
- 存储总空间
- 热数据总空间
- 温数据总空间
- 冷数据总空间
- 文件总数
- 子文件夹总数
- 重复文件总数
- 数据增长趋势
- 文件类型空间占比
- 数据分类空间占比
-



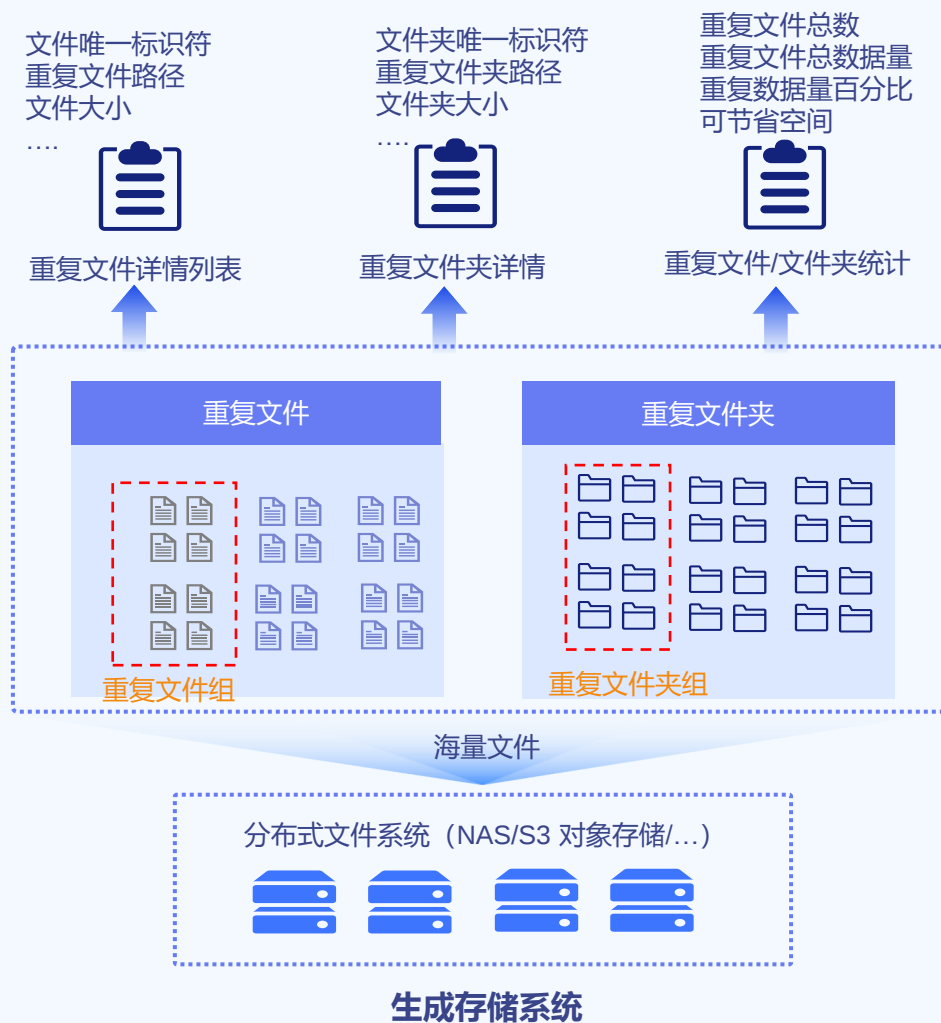
冷数据识别与分析

冷数据分析

- 可筛选需要分析的服务器、最近修改时间、创建时间、最近访问时间、文件大小、一级文件夹
- 冷文件总数
- 冷文件总空间
- 冷数据空间占比
- 不同文件类型冷数据空间占比
- 冷数据分布
- 冷数据空间增长趋势
-



重复文件识别，快速消除冗余数据、降低存储空间



场景描述

在海量文件的存储系统中，常常面临的一个挑战是处理和管理大量的重复数据。重复数据的存在，一方面占用大量存储空间、造成资源浪费，另一方面在备份和恢复过程中会导致备份时间延长和恢复效率降低。同时，重复文件可能存在于企业内的多个存储系统中，如何跨存储识别出海量文件中的重复文件是一大挑战。

价值优势

- **多种重复文件识别方法：**支持文件哈希值抽样比较法、文件属性数据比较法进行重复文件识别。其中文件哈希值抽样比较法通过采样文件数据，实现数据层面的文件重复识别，准确率高。
- **跨异构存储识别重复文件识别，提升管理效率：**通过提供跨异构存储重复文件的识别，存储管理员可针对多个存储系统删除或合并重复文件，提升跨异构存储的管理效率。
- **消除冗余数据、降低存储成本：**通过识别出的重复文件/文件夹详情列表，通过删除或合并重复文件消除冗余数据，可以节省大量的存储空间，并降低存储成本。

释放存储空间

自动扫描存储系统，**识别、合并或删除重复文件，减少不必要的冗余数据。**

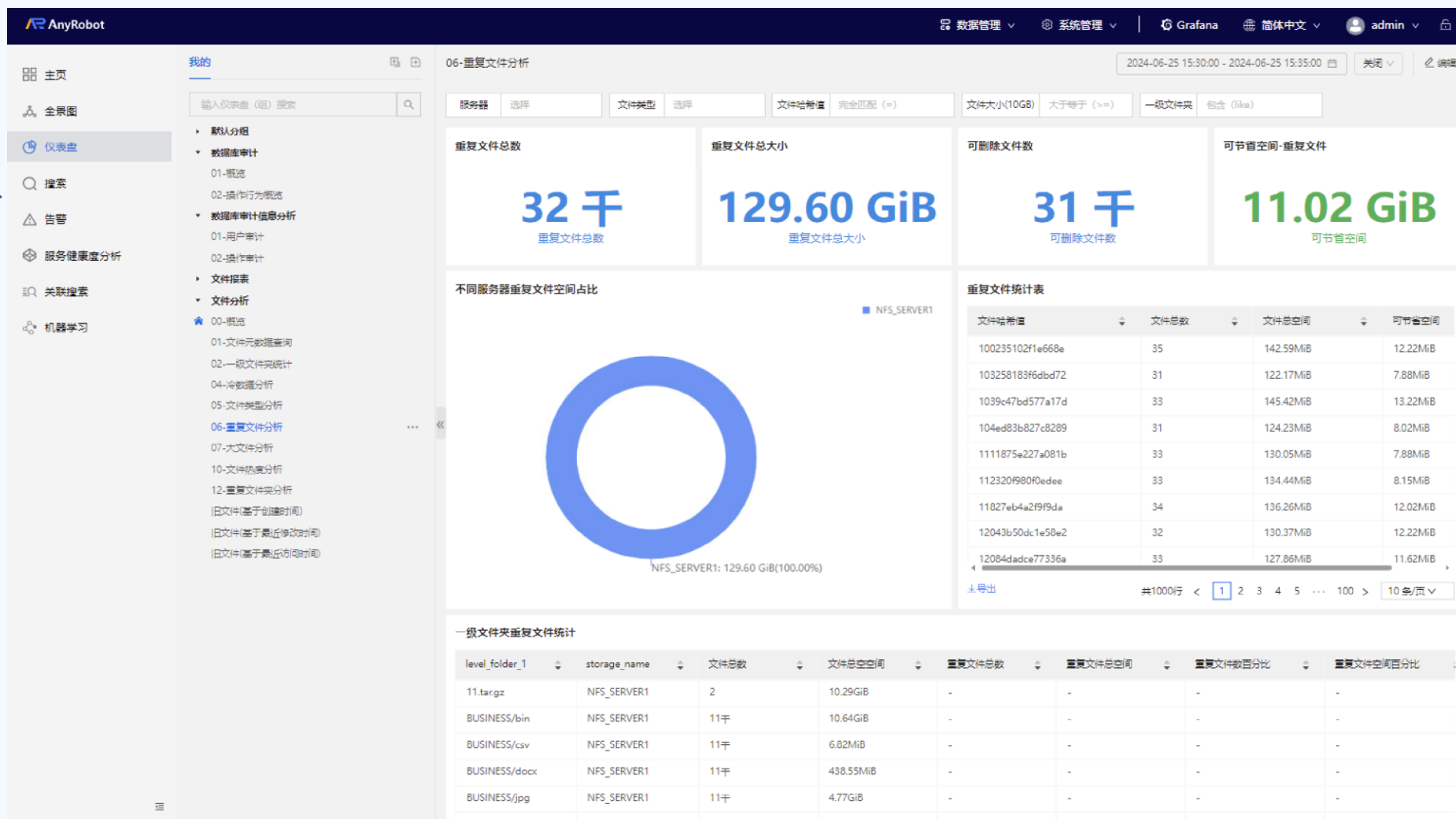
降低管理成本

自动化的文件识别**减少了手动管理和监督的需要，降低由于人为操作导致的错误，从而降低额外的管理成本。**

重复文件分析

重复文件分析

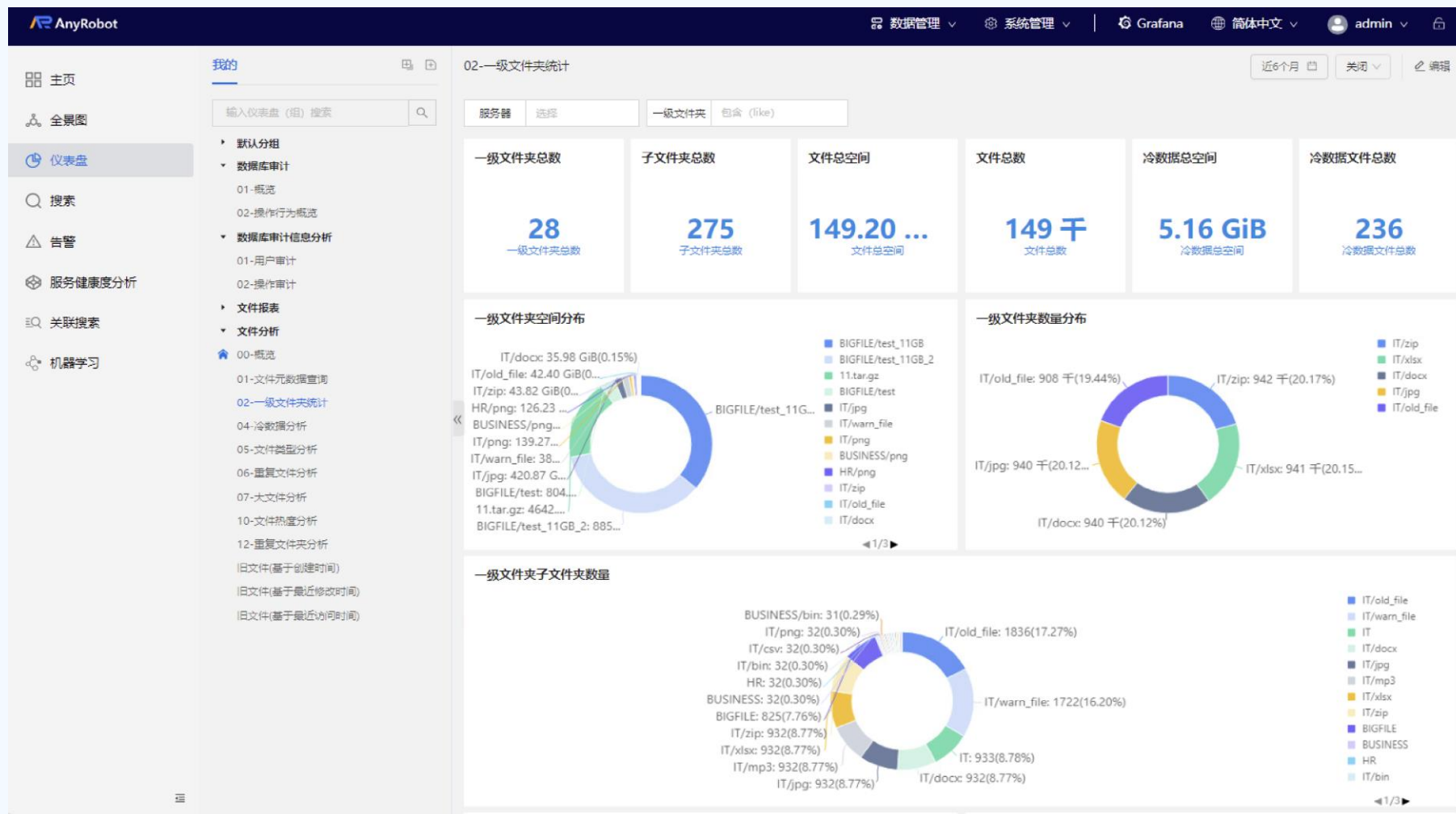
- 可筛选需要分析的服务器、文件类型、文件哈希值、文件大小、一级文件夹
- 重复文件总数
- 重复文件总大小
- 可删除文件数
- 可节省空间
- 重复文件空间占比
- 重复文件统计表
- 一级文件夹重复文件统计
-



重复文件夹统计分析

一级文件夹分析

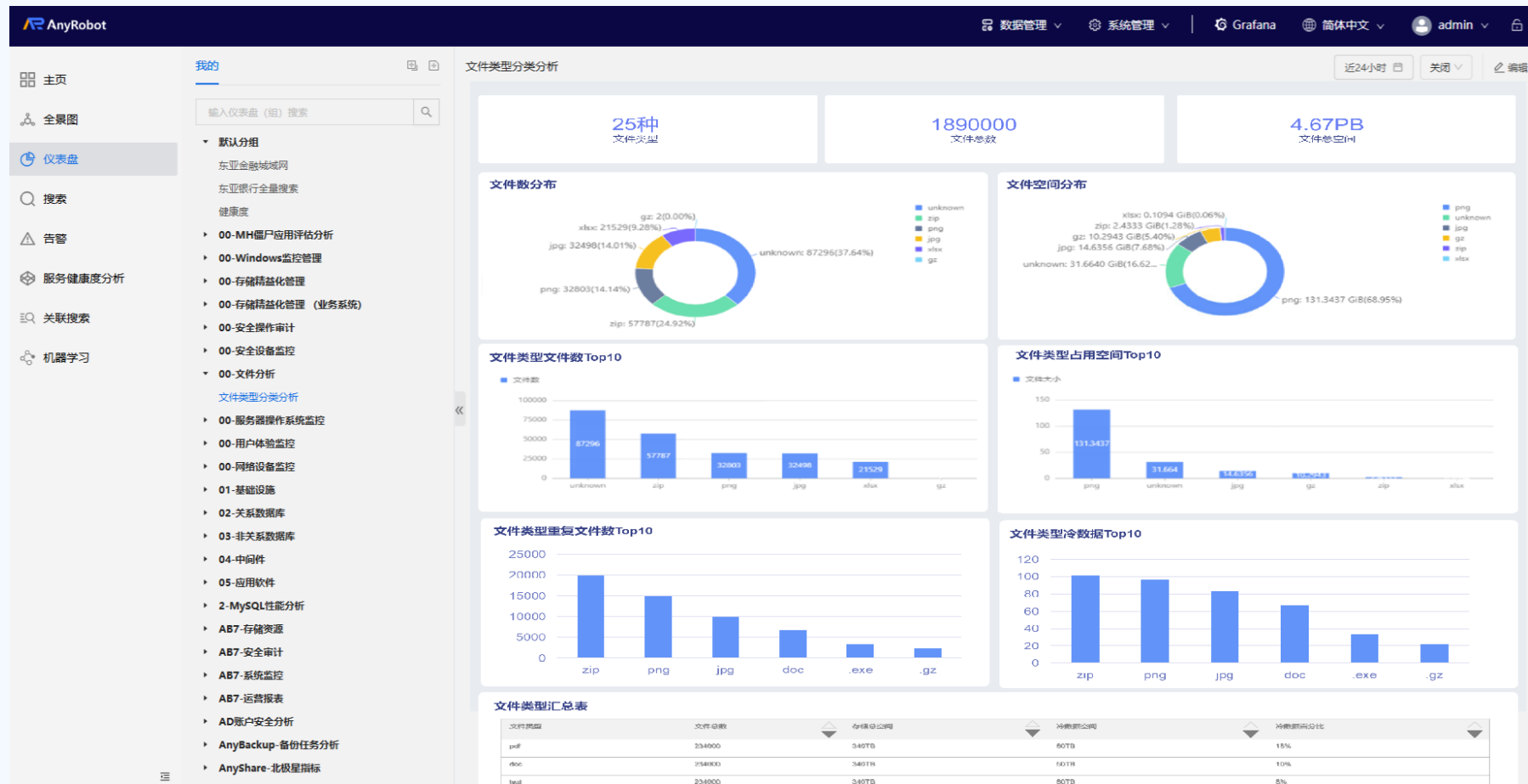
- 可筛选需要分析的服务器、一级文件夹
- 一级文件夹总数
- 子文件夹总数
- 文件总空间
- 文件总数
- 冷数据总空间
- 冷数据文件总数
- 一级文件夹空间分布
- 一级文件夹数据分布
- 一级文件夹子文件夹数量
-



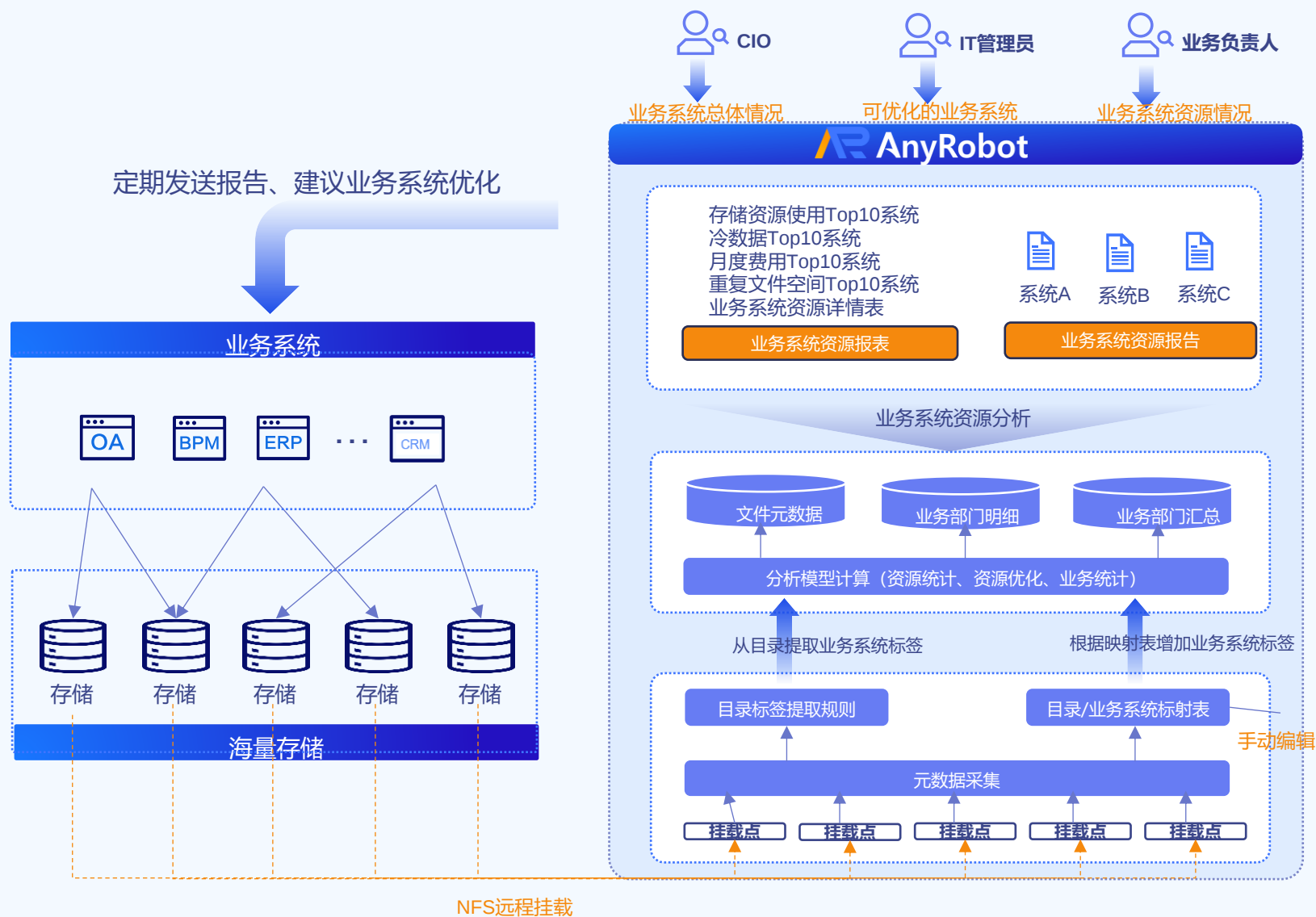
文件类型分类分析

文件类型分类分析

- 文件类型
- 文件总数
- 文件总空间
- 文件数分布
- 文件空间分布
- 文件类型文件数 TOP 10
- 文件类型占用空间 TOP 10
- 文件类型重复文件数 TOP 10
- 文件类型冷数据 TOP 10
- 文件类型汇总表
-



业务系统对存储资源的使用分析和精益化管理



元数据打标

支持两种在文件元数据中打业务部门标签方式

- 目录标签自动提取:** 如果存储中的文件目录带有业务系统信息, 则可通过配置提取规则, 自动提取业务系统标签到元数据
- 目录/业务系统映射表:** 管理员可手动编辑文件目录和业务系统的映射关系, 元数据处理时会根据映射关系附加业务系统标签

业务价值

- CIO:** CIO可从全局了解个业务系统的资源整体使用、可优化的情况
- IT管理员:** IT管理员可了解哪些业务系统资源使用多、可优化空间多, 针对性进行资源优化
- 业务部门负责人:** 可了解负责的业务系统存储资源使用、费用和可优化点, 针对系统存储进行优化

业务系统概览

业务系统存储概览

- 业务系统概览
- 业务系统空间增长趋势
- 优化建议
- 业务系统存储空间分布统计
- 业务系统文件数分布统计
- 业务系统存储空间 TOP 10
- 业务系统费用成本 TOP 10
-



业务系统资源详情

业务系统资源详情

- 可筛选需要分析的业务系统
- 系统资源统计
- 系统优化建议
- 冷数据分布
- 重复文件列表
- 文件类型数据统计
- 文件类型空间统计
- 系统存储空间增长趋势
-



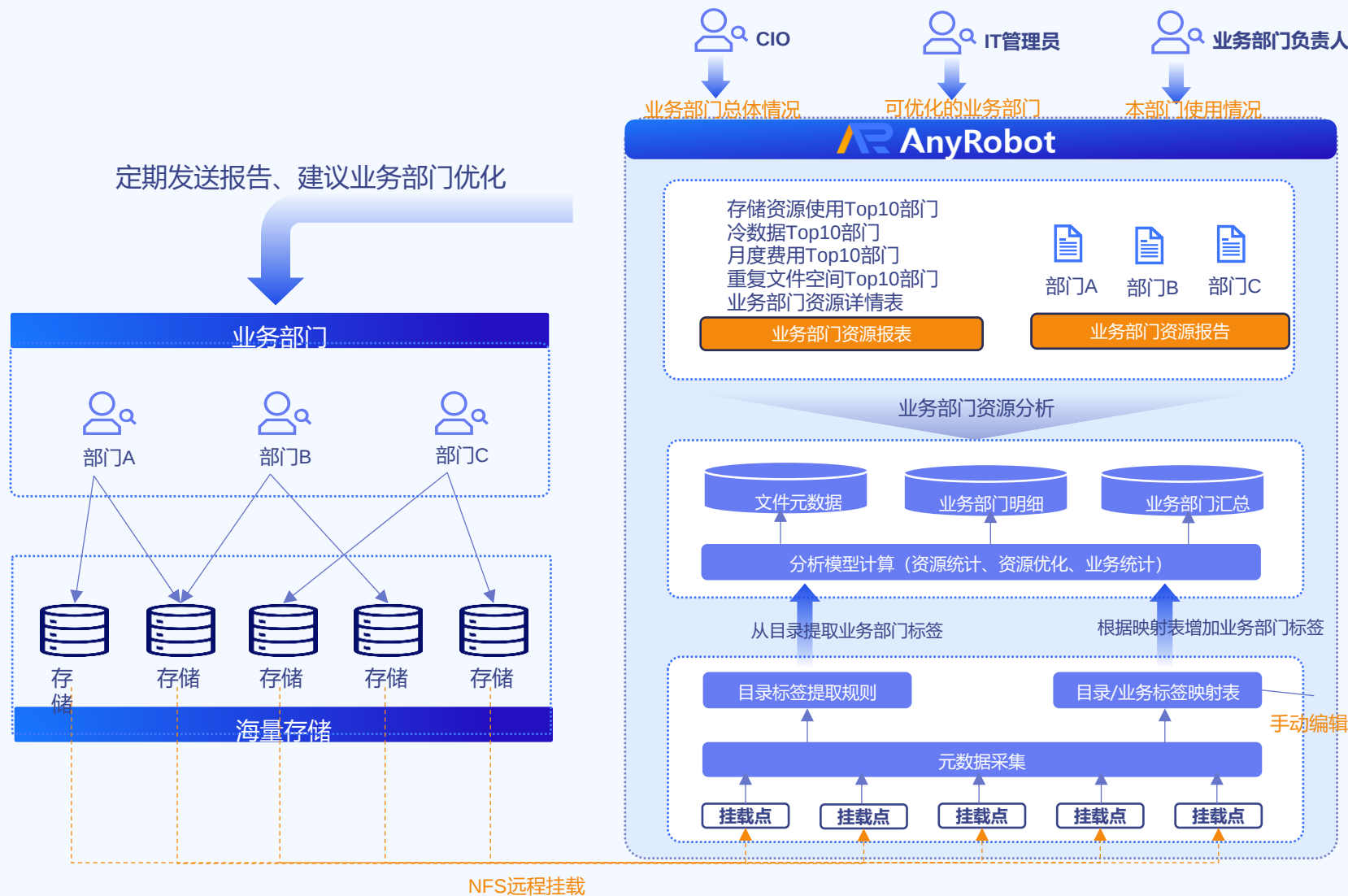
业务系统汇总表

业务系统汇总分析

- 冷数据 TOPN10 业务系统
- 重复文件 TOP10 业务系统
- 业务系统汇总表
-



分析业务部门对存储资源使用现状，实现业务部门成本监控和优化



元数据打标签

支持两种在文件元数据中打业务部门标签方式

- 目录标签自动提取:** 如果存储中的文件目录带有业务部门信息，则可通过配置提取规则，自动提取业务部门标签到元数据
- 目录/业务部门映射表:** 管理员可手动编辑文件目录和业务部门的映射关系，元数据处理时会根据映射关系附加业务部门标签

业务价值

- CIO:** CIO可从全局了解个业务部门的资源整体使用、可优化的情况
- IT管理员:** IT管理员可了解哪些业务部门资源使用多、可优化空间多，针对性进行资源优化
- 业务部门负责人:** 可了解本部门资源使用、费用和可优化点，针对部门存储进行优化

业务部门概览

业务部门概览

- 业务使用概览
- 业务部门空间增长趋势
- 优化建议
- 存储空间业务部门分布统计
- 成本费用业务部门分布统计
- 业务部门存储空间 TOP 10
- 业务部分费用成本 TOP 10
-



业务部门资源详情

业务部门资源详情分析

- 部门使用统计
- 部门优化建议
- 冷数据分布
- 重复文件列表
- 文件类型数据统计
- 文件类型空间统计
- 部门空间增长趋势
-



业务部门汇总表

业务部门汇总分析

- 冷数据 TOP 10部门
- 重复文件 TOP 10 部门
- 业务部门汇总表
-



PART 2.2

提高IT运维效率

- 跨平台海量文件检索
- 存储容量预测
- 自动化报告

跨平台文件检索，全面洞察文件存储大小和分布



全面的数据视图

- 全面洞察不同位置和系统中的文件数据分布，有效地整合分散的数据，构建统一的数据视图。

多视角数据资源清点

- 业务视角**：识别不同业务部门的数据需求，并根据业务标准对文件进行分类和标记，从业务视角梳理资源大小与分布；
- 个人使用者视角**：分析员工在虚拟桌面等企业资产中的文件使用行为，盘点个人使用者的数据资源大小与分布；
- 自定义视角**：支持基于特定需求自定义洞察角度，如按项目、部门等。

跨平台文件路径检索，合理规划存储使用

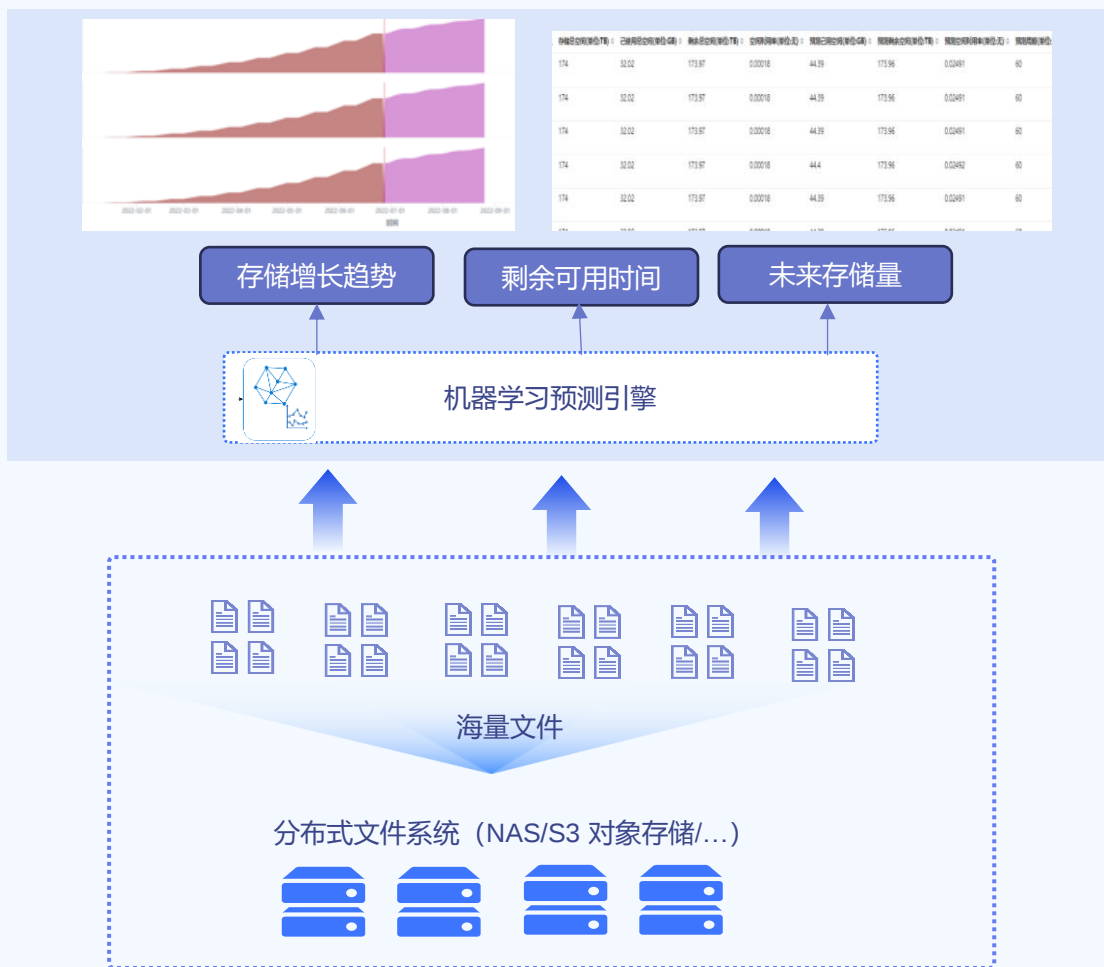
文件元数据查询

- 可筛选需要分析的服务器、文件类型、文件热度、创建时间、最近修改时间、最近访问时间、元数据哈希值、文件哈希值、文件路径、一级文件夹、文件名
- 所有文件详情列表
-

The screenshot displays the 'AnyRobot' File Metadata Query interface. The sidebar on the left contains navigation links: '我的' (My), '仪表盘' (Dashboard), '搜索' (Search), '告警' (Alerts), '服务健康度分析' (Service Health Analysis), '关联搜索' (Associated Search), and '机器学习' (Machine Learning). The main content area is titled '01-文件元数据查询' (01-File Metadata Query) and includes a search bar and various filters. The filters include '服务器' (Server), '文件类型' (File Type), '文件大小' (File Size), '创建时间' (Creation Time), '最近修改时间' (Last Modified Time), '最近访问时间' (Last Access Time), '元数据哈希值' (Metadata Hash Value), '文件哈希值' (File Hash Value), '文件路径' (File Path), and '一级文件夹' (First-level Folder). The table below shows a list of files with columns for '时间' (Time), '文件路径' (File Path), '文件名' (File Name), '服务器' (Server), and '文件类型' (File Type).

时间	文件路径	文件名	服务器	文件类型
2024-06-27 00:00:06.200	/root/NFS_SERVER1/IT/jpg/jpegkaml_pi.jpg	jpegkaml_pi.jpg	NFS_SERVER1	jpg
2024-06-27 00:00:06.200	/root/NFS_SERVER1/IT/jpg/jpegzumamwfq.jpg	jpegzumamwfq.jpg	NFS_SERVER1	jpg
2024-06-27 00:00:06.200	/root/NFS_SERVER1/IT/warn_file/jpg/jpegx6_1x8xw.jpg	jpegx6_1x8xw.jpg	NFS_SERVER1	jpg
2024-06-27 00:00:06.199	/root/NFS_SERVER1/IT/jpg/jpegypgoxlrd.jpg	jpegypgoxlrd.jpg	NFS_SERVER1	jpg
2024-06-27 00:00:06.199	/root/NFS_SERVER1/IT/jpg/jpegyf56x6xg.jpg	jpegyf56x6xg.jpg	NFS_SERVER1	jpg
2024-06-27 00:00:06.199	/root/NFS_SERVER1/IT/jpg/jpegyetnpv32.jpg	jpegyetnpv32.jpg	NFS_SERVER1	jpg
2024-06-27 00:00:06.199	/root/NFS_SERVER1/IT/jpg/jpegzk5hcat7.jpg	jpegzk5hcat7.jpg	NFS_SERVER1	jpg
2024-06-27 00:00:06.199	/root/NFS_SERVER1/IT/jpg/jpegzqrwk9l.jpg	jpegzqrwk9l.jpg	NFS_SERVER1	jpg
2024-06-27 00:00:06.199	/root/NFS_SERVER1/IT/jpg/jpegyewd97nd.jpg	jpegyewd97nd.jpg	NFS_SERVER1	jpg
2024-06-27 00:00:06.199	/root/NFS_SERVER1/IT/jpg/jpegys59cam6.jpg	jpegys59cam6.jpg	NFS_SERVER1	jpg

存储容量预测，降低容量规划复杂度



- **深度学习:** TCN、RNN、block_RNN、nbeats、transformer
- **ARIMA:** ARIMA、Auto_ARIMA
- **基线算法:** naïve_arift、naïve_drift、naïve_mean、naïve_seasonal
- **指数平滑:** SES、Holt

前瞻性存储资源分配

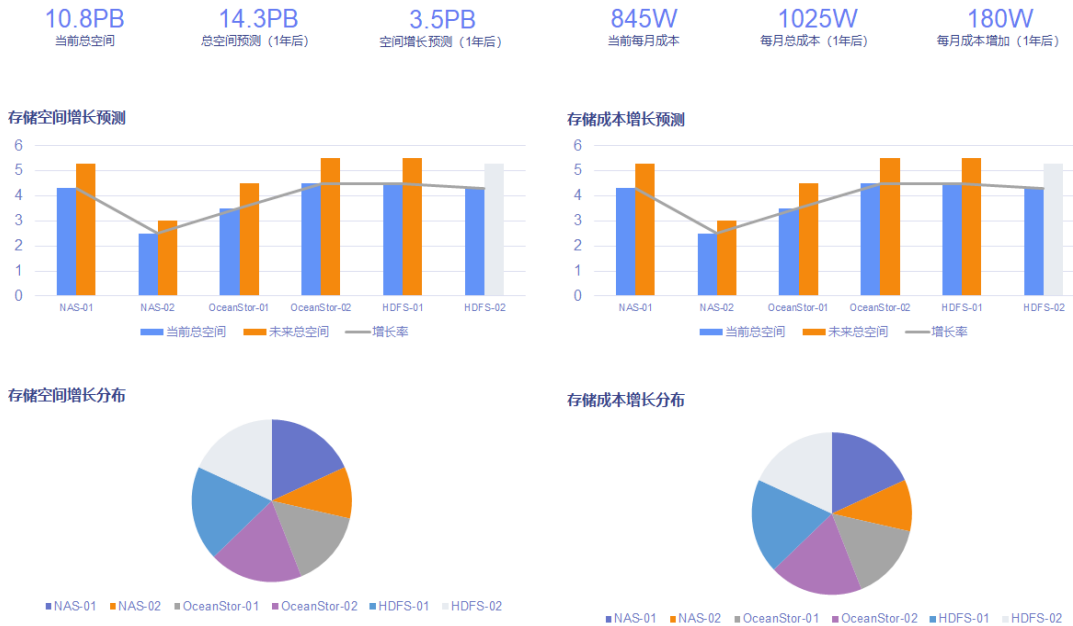
- 预测性规划使企业能够根据预期的存储增长趋势提前为不同的业务部门分配存储资源，避免存储空间不足。

避免资源瓶颈

- 预测存储的剩余使用时间、未来存储量可以及时调整存储资源配置，防止因存储容量不足导致的数据丢失等问题。

存储容量预测

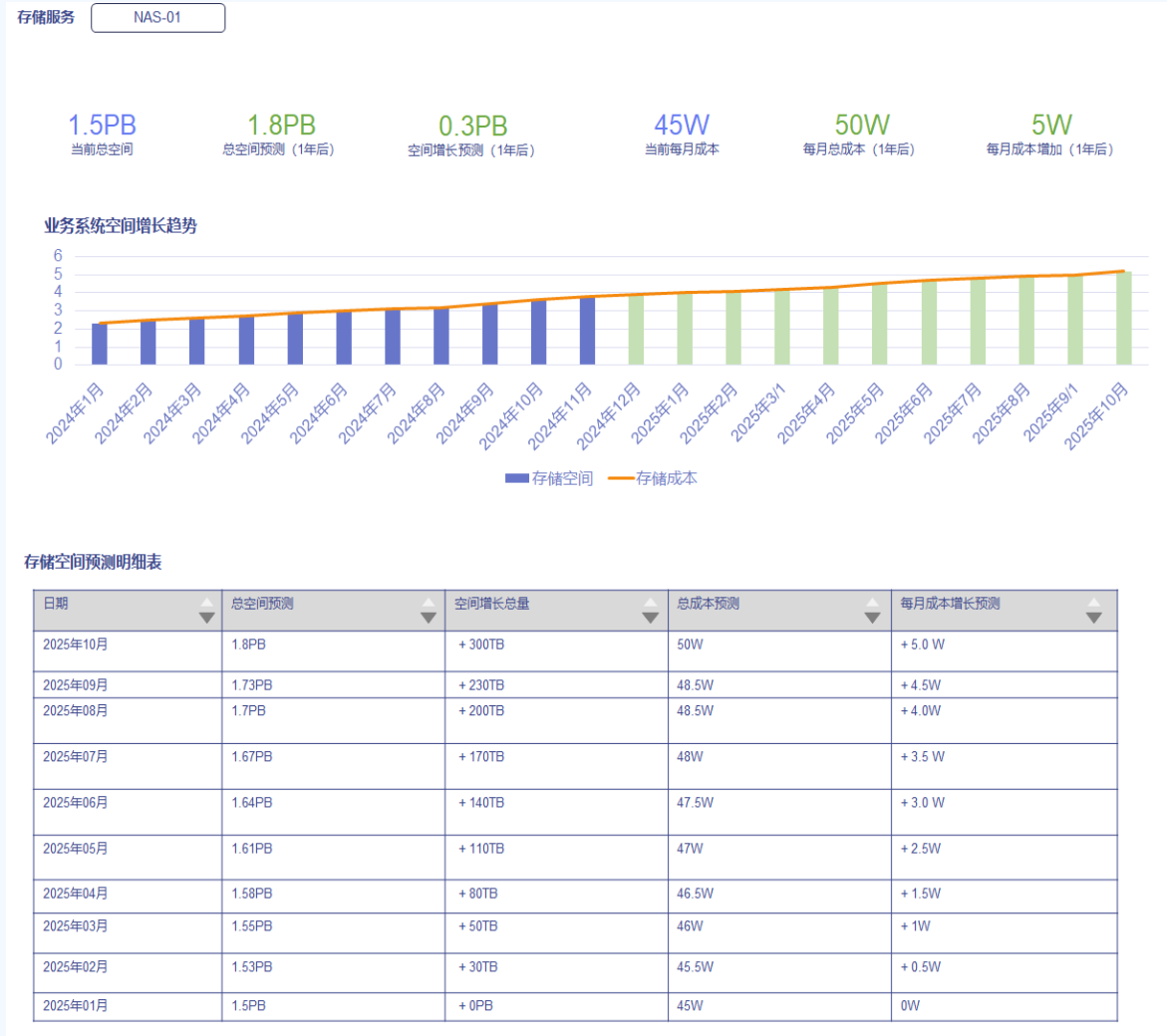
存储容量预测概览



存储空间预测明细表

存储服务	存储总空间	文件/总数	总空间预测 (1年)	空间增长预测 (1年)	总成本预测 (1年)	成本增长预测 (1年)
NAS-01	600TB	300000000	800TB	+200TB	1450000元	+ 250000
NAS-02	800TB	300000000	1000TB	+200TB	1650000元	+ 250000
OceanStor-01	1.5PB	300000000	1.8PB	+300TB	4550000元	+ 300000
OceanStor-02	2.3PB	300000000	2.6PB	+300TB	5450000元	+ 300000
HDFS-01	4.2PB	300000000	5PB	+800TB	6450000元	+ 300000
HDFS-02	2.5PB	300000000	2.8PB	+300TB	4450000元	+ 300000

存储容量预测详情



自动化报告存储资源使用，提升管理效率



场景描述

运维管理员需要定期向领导以及业务部门出具存储系统的报告，包括存储系统总量、增长趋势、各业务部门使用情况、可优化的空间等。同时希望减少手动生成报告的时间和人力成本，通过自动化工具快速生成准确、一致的报告。企业希望不同的部门或用户可能对报告有不同的需求，需要能够自定义报告的内容、格式和呈现方式

业务价值

- **为领导决策提供依据**：通过向领导存储系统整体报告，为领导的决策提供有力支持
- **促进业务部门优化**：业务部门报告，帮助业务部门了解本部门的增长趋势、可清理数据以及数据资产情况，促进业务部门优化存储。
- **降低手动成本、提升效率**：定期生成自动化报告，无需手动收集数据、生成报告，降低人工成本、提升效率。

自定义报告模版

支持依据业务、个人使用者或其他特定需求和偏好自定义数据分类，并依据分类**设计报告模板**。

定期报告生成

支持**设置定期生成报告的规则**，无论是每日、每周还是每月，都能准时产出所需的报告，并以邮件方式发送给指定用户。

存储设备性能监控

存储设备性能监控

- 各分区空间使用率
- 磁盘I/O使用率
- 磁盘读写速率
- 网络
- CPU
- 内存
-



存储精益化管理解决方案价值总结

降低存储成本

- 识别不常访问的冷数据，并将这些数据迁移到成本更低的存储介质上，**减少对高性能但高成本存储资源的依赖。**
- 自动识别重复或冗余的文件，帮助客户释放存储空间，**减少不必要的存储开销。**
- 提供基于数据管理方式和业务需求的存储优化建议，帮助用户**实现更高效的存储资源利用。**

提高IT运维效率

- 快速检索和分析存储在不同平台和位置的文件，从全局视角、特定的业务视角、个人使用者视角或其他自定义视角洞察数据资源分布和大小，**提高数据资源的管理效率。**
- 利用机器学习预测引擎预测未来的存储需求，**提前规划资源，避免存储容量不足的问题。**
- 自动化报告**减少了手动管理和监控存储资源的工作量**，降低管理成本的同时，提高了IT运维效率。

PART 03

工作实施内容及优化成效

存储资源盘活工作开展

工作内容	预期效果	责任人
调研现有数据分级情况	了解现有数据类型、数据分级、存储管理规范	爱数顾问主导、行方存储负责人配合，行方需协助协调
现状详细评估	针对现有数据分类评估优化方向	爱数顾问主导、行方存储负责人配合，行方需协助协调
挑选10套试点业务系统（数量可调整），利用AnyRobot工具进行分析验证	自动化采集并形成数据热温冷分析结果和重复文件识别结果，并进行人工确认及修正	爱数顾问主导、行方存储负责人配合，行方需协助协调
根据分析效果形成下一步优化计划与联创计划，落实执行	考虑到银行业务的行业属性，在指标与模型方面需要有针对性地优化，所以明确优化目标，并共同联创执行，后续可共同申请专利与奖项	爱数顾问与研发主导、行方存储负责人配合，行方需协助协调
存储数据资产视图展示与完善	存储数据资产全局视角管理、存储使用分析、业务部门使用分析、个人使用分析、自定义分析	爱数顾问主导、行方存储负责人配合，行方需协助协调
配套制度的优化与完善	制定出贴近于银行的数据分类管理制度，便于与工具结合有效落实执行	爱数顾问主导、行方存储负责人配合，行方需协助协调

某企业存储优化成效

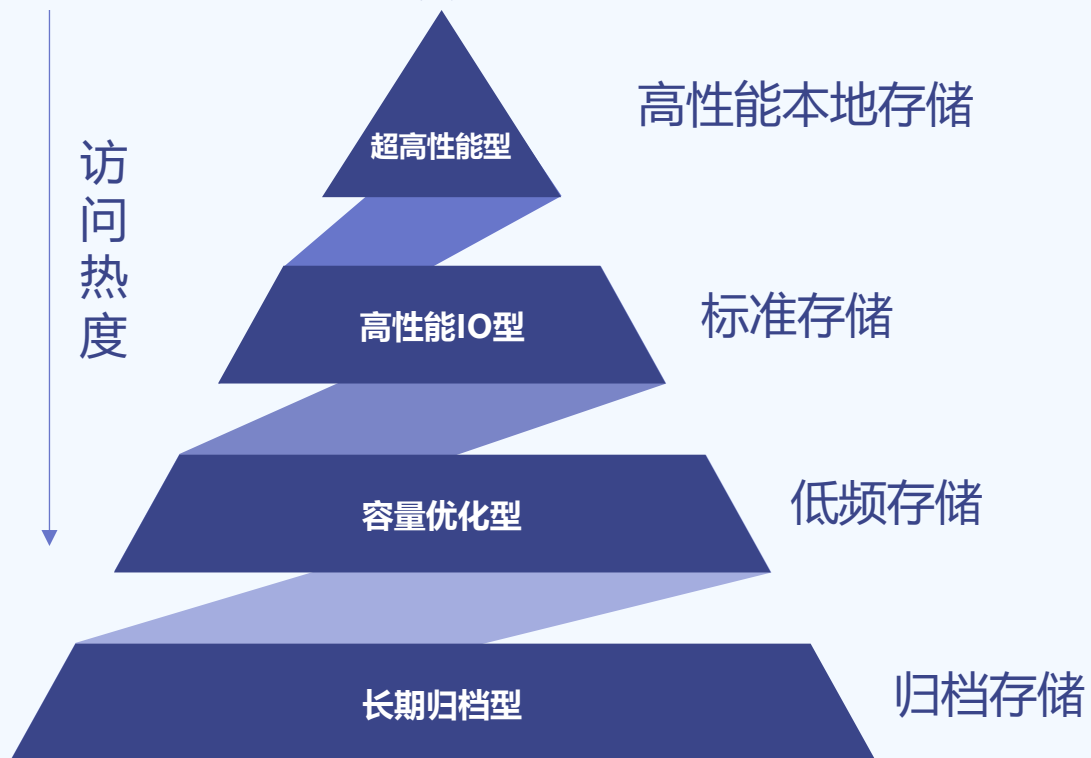
优化前:

- 某企业每年存储采购**1200万**左右, 高端存储占比**65%**
- 存储数据情况不清楚
- 存储资源管理混乱



优化后: 节省约**492万**直接投入

- 每年节省高端存储采购约312万左右
($1200 \times 0.65 \times 0.4 = 312$)
- 每年重复文件识别节省约180万左右
($1200 \times 0.15 = 180$)



管理内容	优化收益	实现成效
冷数据识别	识别冷数据, 并将这些数据迁移到成本更低的存储介质上	冷数据识别迁移节省约 40% 存储空间
重复文件识别	自动识别重复或冗余的文件, 帮助客户释放存储空间	重复文件识别率95%以上, 节省约 15% 存储空间
存储容量预测	提前规划资源, 避免存储容量不足的问题	一年减少存储资源不足造成的生产事件 2 次
存储资源管理	减少人工查询、汇总和报告的时间	减少 30% 的人工时间