HEPS科学数据管理

胡 皓，罗齐

HEPS,CC / IHEP,CC

2020.10.19

科学数据管理目标与任务

- 数据管理标准与规范：国家规范+领域规范+用户认可...
- 元数据获取：在实验不同阶段从各系统（控制、用户服务、数据分析）获取与数据相关的元数据
- 数据索引：实现对原始数据和衍生数据的catalogue
- 数据存储：数据标准化、数据长期保存
- 数据服务：数据下载、访问、利用（包括分析）与追踪
- 开放接口：标准接口，满足各系统、各模块协作与通讯

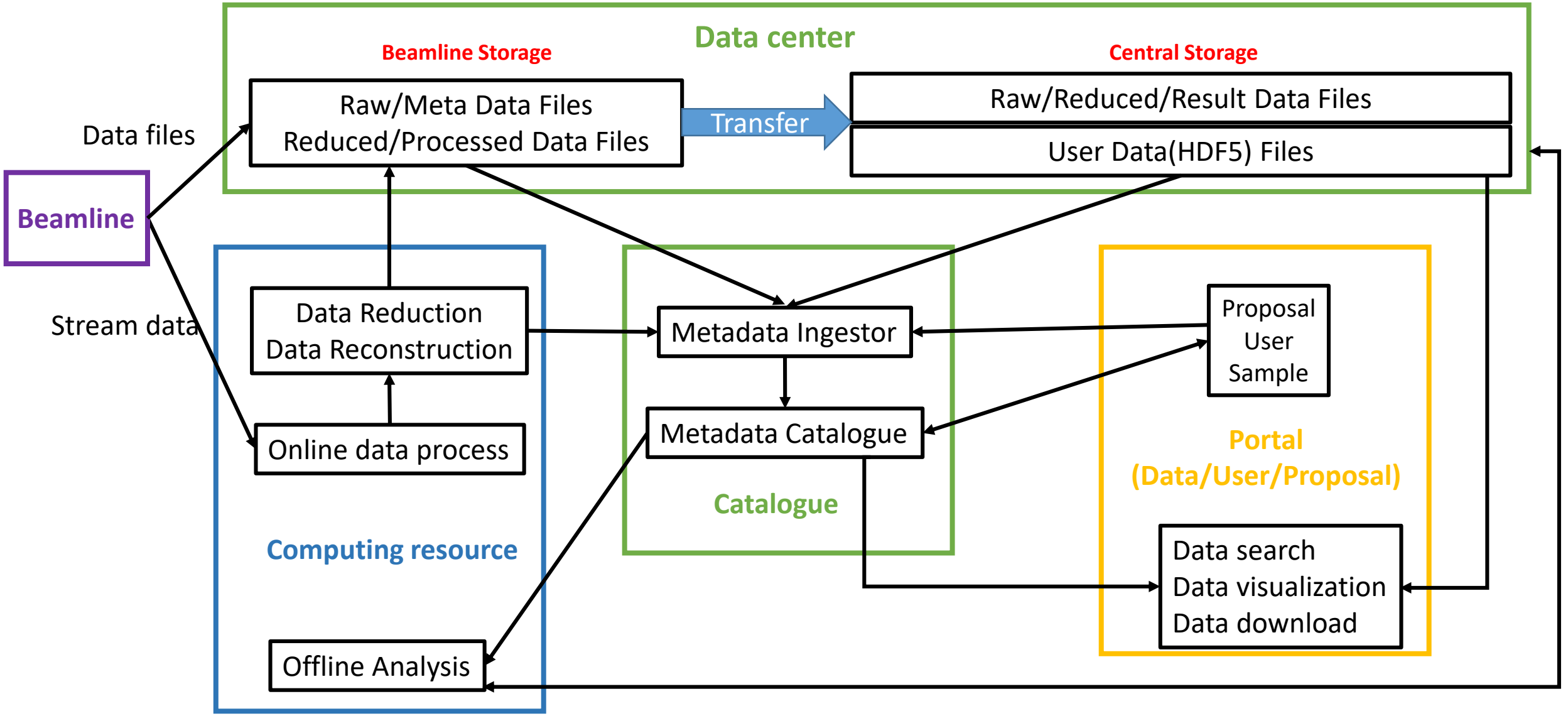
数据策略

- 建议数据长期保存
- 建议提供至少三个月磁盘存储
- 建议提供永久磁带存储
- 对raw data、result data、 calibration data, user data长期保存
- 对reduced data、processed data短期保存
- 支持数据开放共享：数据封闭期设置
- 每个数据集都有唯一永久标识（PID）
- 任何与该数据集相关的文章必须引用数据集的PID

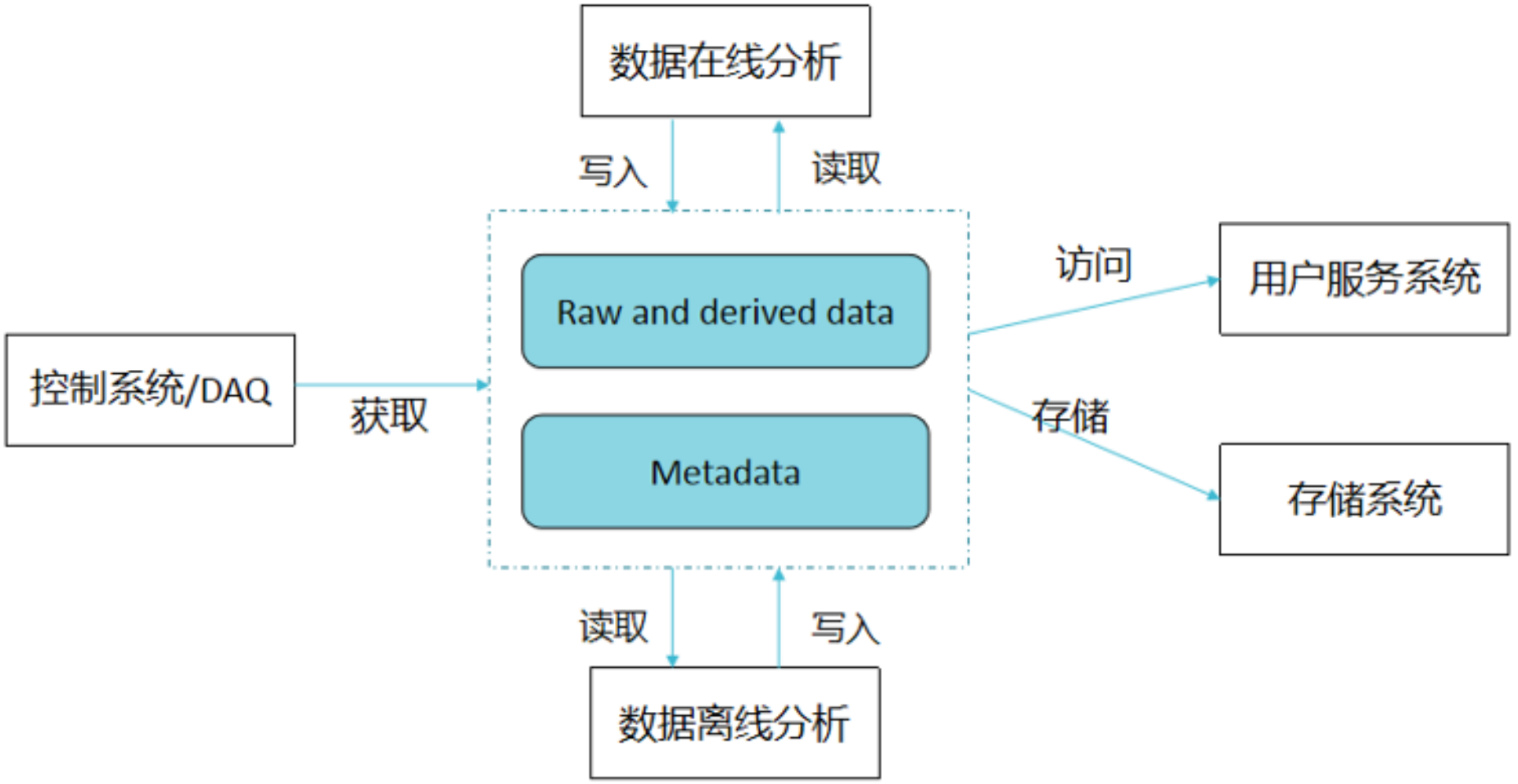
数据策略初稿已经完成，细节需要再讨论

参考PaN data Europe 数据策略http://pan-data.eu/sites/pan-data.eu/files/PaN-data-D2-1.pdf

数据流


Data center
Beamline Storage
Central Storage
Raw/Meta Data Files
Reduced/Processed Data Files
Transfer
Raw/Reduced/Result Data Files
User Data(HDF5) Files
Data files
Beamline
Stream data
Data Reduction
Data Reconstruction
Online data process
Computing resource
Metadata Ingestor
Metadata Catalogue
Catalogue
Proposal
User
Sample
Portal
(Data/User/Proposal)
Data search
Data visualization
Data download
Offline Analysis

数据管理系统接口


数据在线分析
写入
读取
Raw and derived data
Metadata
控制系统/DAQ
获取
访问
用户服务系统
存储
存储系统
读取
写入
数据离线分析

数据文件存储策略和存储结构

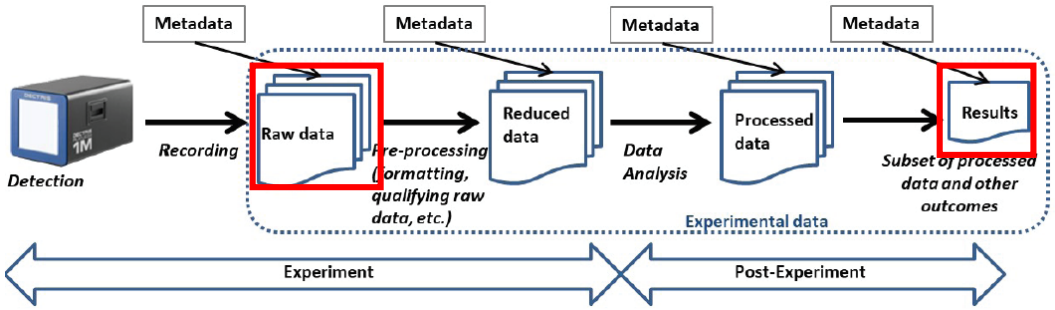
- 数据存储策略：

控制系统/DAQ → Raw data:长期存储

分析系统 → Reduced data/Processed data:短期保存

分析系统→Result data:长期存储

User data:以统一标准格式长期存储



- 存储目录结构设计

Beamline storage

Central storage

```
|-- <Beamline>
    |-- <Year>
    |    |-- <data>
    |    |    |-- <beamtime-ID>
    |    |    |    |-- <raw>  Read only
    |    |    |    |    |-- <samplename>
    |    |    |    |    |    |-- <scanID>
    |    |    |    |-- <processed> Read/Write
    |    |    |    |-- <scratch> Read/Write
    |    |-- <commissioning>
    |    |    |-- <commissioning-ID>
    |    |    |    |-- <raw>  Read only
    |    |    |    |-- <processed> Read/Write
    |    |    |    |-- <scratch> Read/Write
|-- common
    |-- B01
    |-- B02
    |-- ...
    |-- B14
```

beamtime-ID：
用户每次被批复机时会分配一个唯一的ID，对应唯一的Proposal ID

commissioning-ID：
测试机时ID，对应调试环境实验数据

每个线站刻度数据

元数据来源及分类

- **科学/实验元数据(scientific metadata)**

  What: 样本信息,实验环境参数

  From: 控制系统

  For: 数据分析 & 元数据目录管理

  To: HDF5文件 , 元数据库

- **管理元数据(administrative metadata)**

  What: 提案信息,用户信息,实验类型,线站等
  数据文件信息（PID，存储路径，大小）

  From: 提案系统、用户系统，DAQ/控制系统

  For: 元数据目录管理

  To: 元数据库

- **其他元数据(other metadata)**

  What: 数据状态、传输状态、传输校验信息、
  分析软件,更新时间等

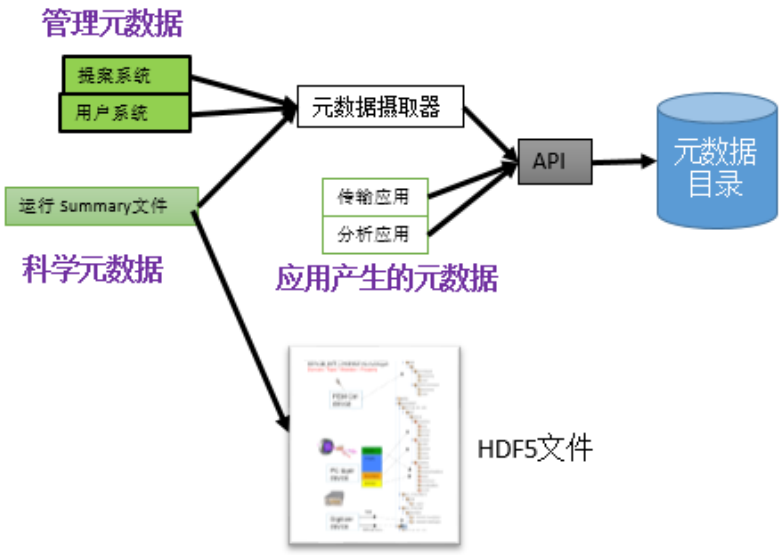
  From: 传输应用,分析应用

  For: 元数据目录管理

  To: 元数据库

ESRF opendata 数据集相关实验元数据

MX_aperture	50 um
MX_beamShape	ellipse
MX_beamSizeAtSampleX	0.05
MX_beamSizeAtSampleY	0.05
MX_dataCollectionId	1719726
MX_detectorDistance	234.913
MX_directory	/data/id30a1/inhouse/opid30a1/20160408/RAW_DATA/AFAMIN/
MX_exposureTime	0.1
MX_flux	141000000000
MX_fluxEnd	141000000000
MX_motors_name	y z sampx sampy phi kappa chi kappa_phi zoom focus phiz phiy
MX_motors_value	2.041 -0.596 0.033 0.233 149.5 0.0 0.0 0.0 2.0 0.0 0.0 -0.605
MX_numberOfImages	24
MX_oscillationOverlap	0
MX_oscillationRange	0.0416667
MX_oscillationStart	149.5
MX_resolution	1.99991
MX_scanType	OSC
MX_startImageNumber	1



Metadata acquisition approaches

Depends on how metadata are provided

- Provide interfaces for control system to write metadata to DMS(recommended)
- Metadata-ingestor plugins are designed to collect metadata from nexus/txt/HDF5 files
- Ingest metadata from DAQ database directly

Metadata acquisition approach（1）--from control system

- Suitable for self-designed detector, having ability to program to control software(Epics/Tango）
- Interfaces should be provided for control system
- With which control system can send metadata (JSON-based) to kafka broker when a dataset is produced.

README.md

Example from github

Area detector Kafka interface

This repository contains two separate projects which facilitates the transmission of data between an EPICS IOC and an Apache Kafka broker. The two projects are:

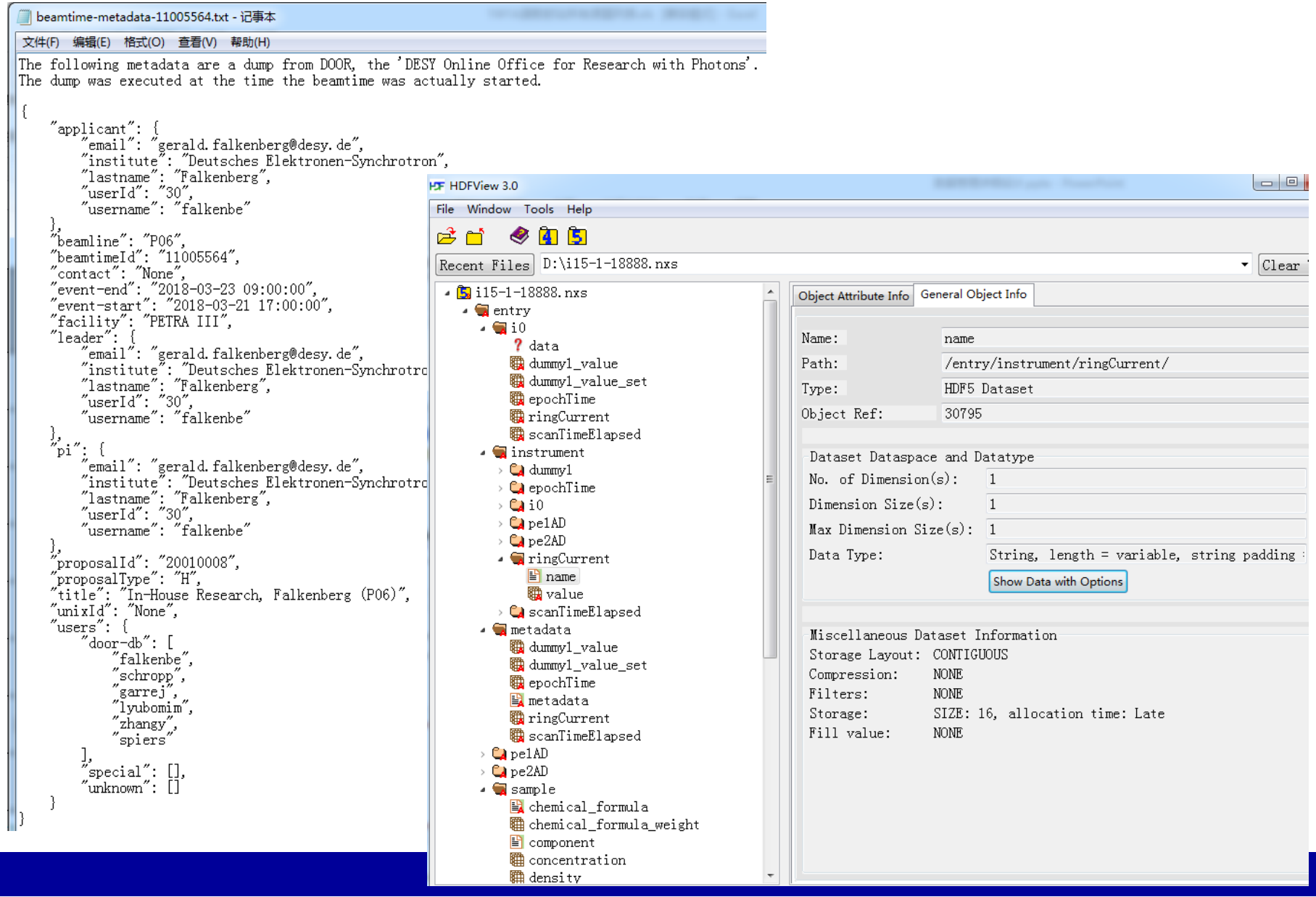
- An EPICS areaDetector driver which acts as a Kafka consumer and makes NDArray data received from the broker available to the IOC.
- An EPICS areaDetector plugin which connects to an areaDetector and serializes NDArray data it receives and sends it to a Kafka broker.

Apache Kafka is an open-source platform for handling streaming data one or more data brokers in order to maximize throughput and reliability. More information on Apache Kafka can be found at the website of that project.

For serializing and de-serializing the areaDetector (NDArray) data, Google FlatBuffers is used. Serializing data using FlatBuffers is fast with a relatively small memory overhead while being easier to use than C-structs.

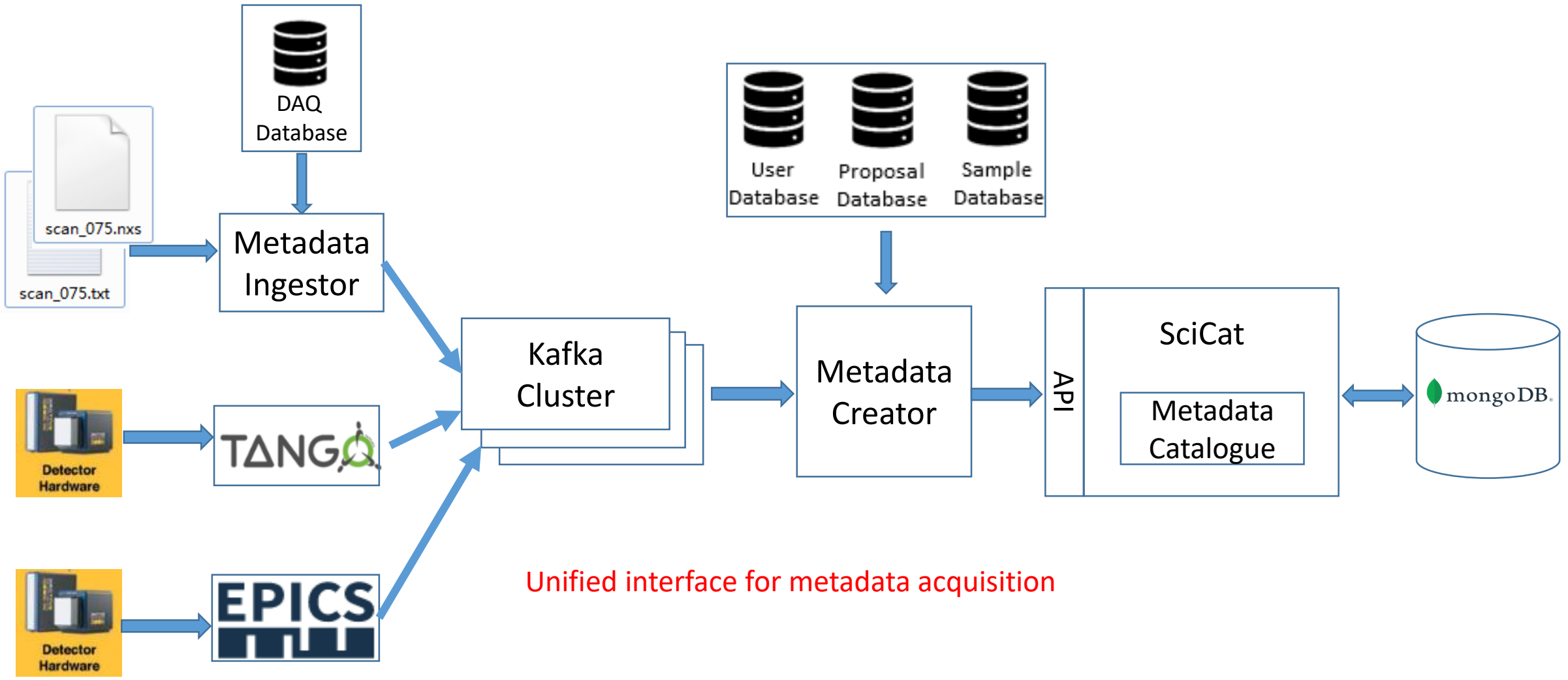
```
{ topic: 'bm01_detector1_topic',
  messages: [{
            "proposalID":"p0001",
            "sampleID":"s0001",
            "userAccount":"wch@ihep.ac.cn",
            "creationLocation":"bm01_detector1"
            "dataset_name":"bm01_1_38883",
            "dataFormat":"hdf5",
            "path":"/nfs/B01/2020/data/73838/raw/scan01",
            "scientificMetadata":{
                        elog_id:"234",
                        optical_coupling:"dryfit"},
            "description":"XXX research"
            }]
 }
```

Metadata acquisition approach（2）--from files

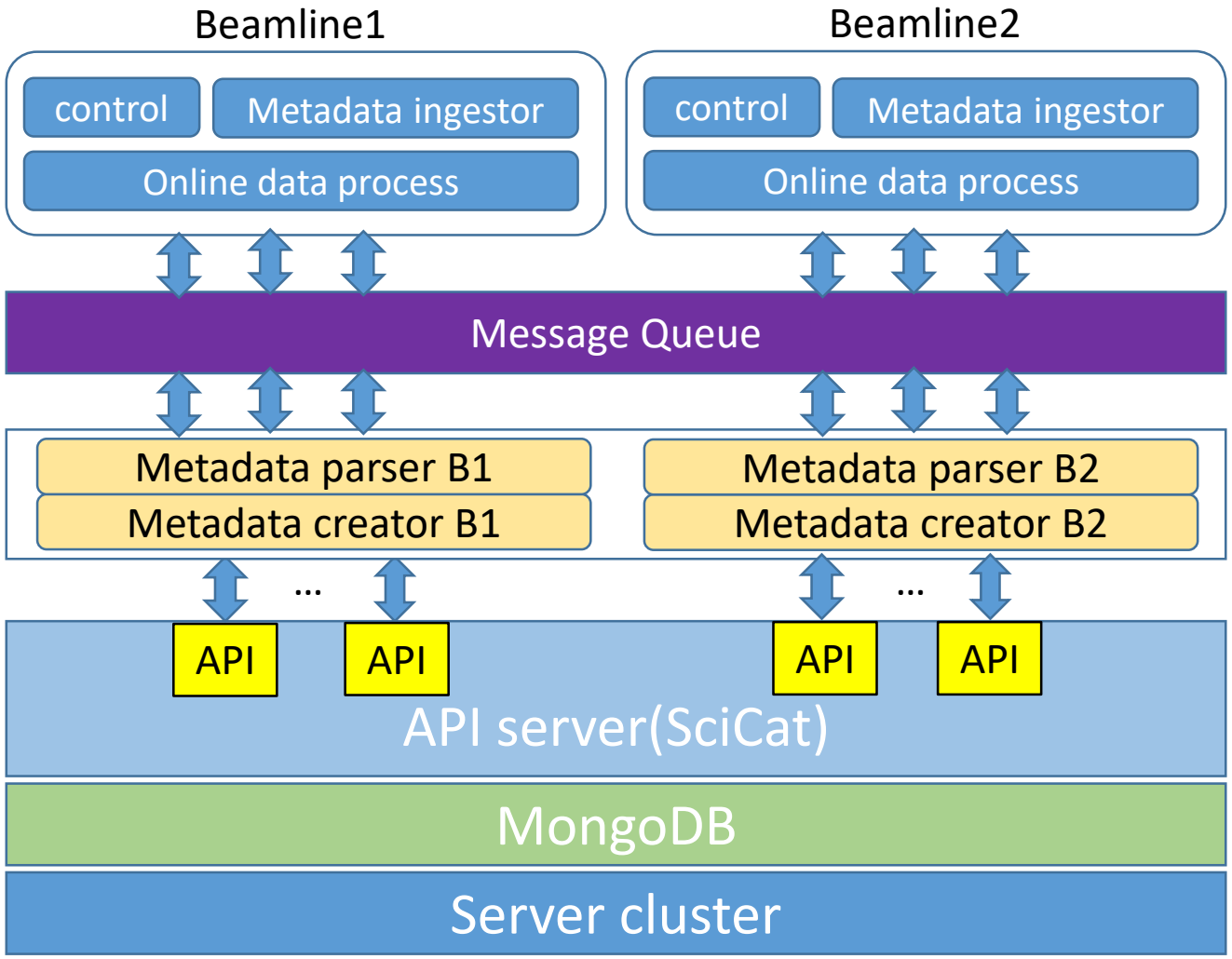


- Suitable for those that metadata are collected and saved to files (txt/nxs)
- Metadata-ingestor runs in flexible plugin mode.
- In addition to user and experiment information, the plug-in also records information such as file size, checksums, processing time, and file replicas.

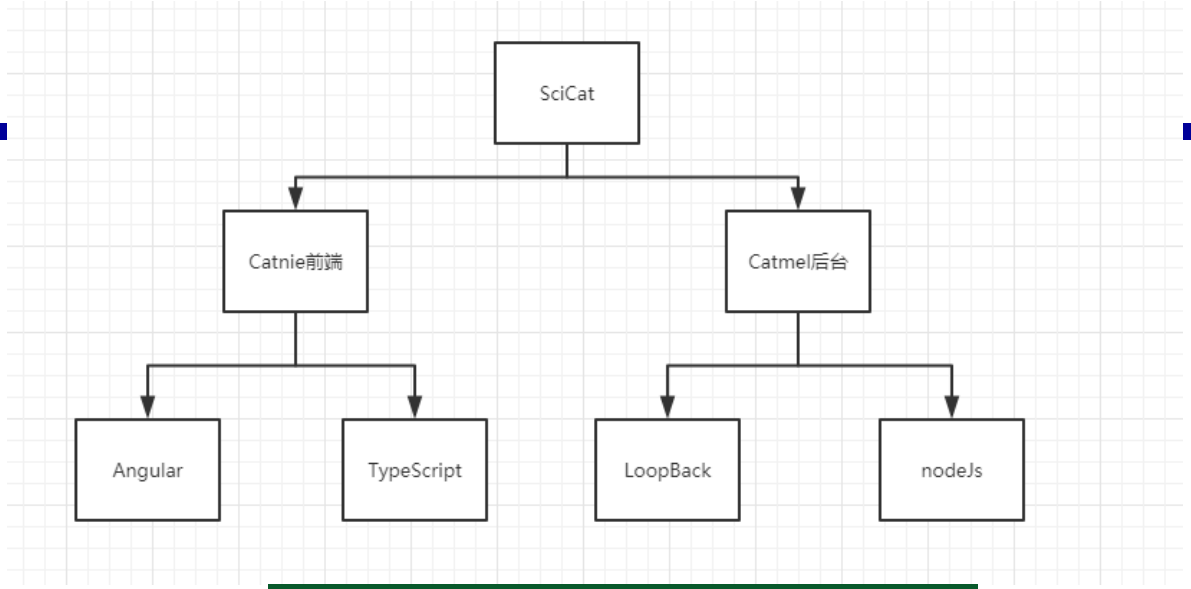
Metadata acquisition— for cataloging



元数据获取部署架构


Beamline1
control
Metadata ingestor
Online data process
Beamline2
control
Metadata ingestor
Online data process
Message Queue
Metadata parser B1
Metadata creator B1
Metadata parser B2
Metadata creator B2
...
...
API
API
API
API
API server(SciCat)
MongoDB
Server cluster

SciCat说明及应用



前端Catanie

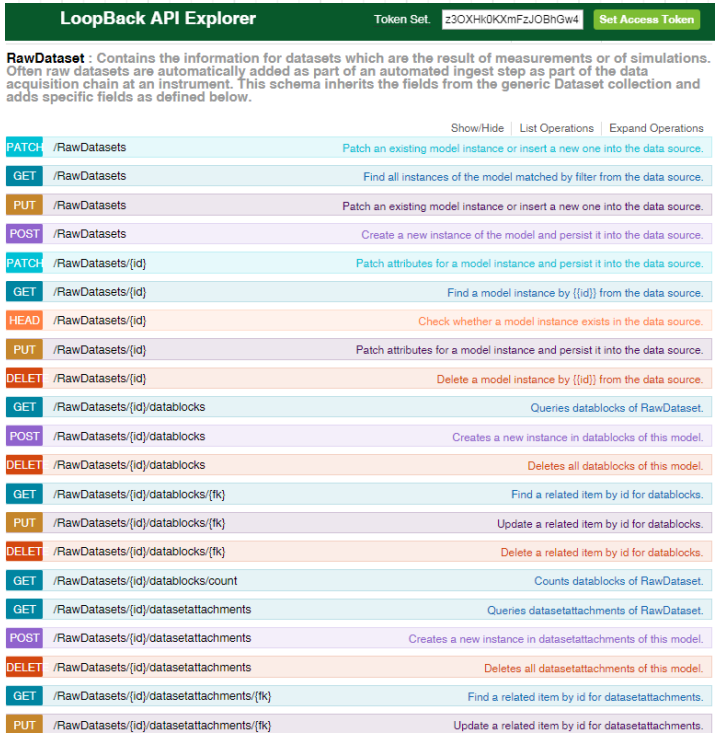
AngularJS：JavaScript 框架。AngularJS主体使用TypeScript编写。

后端Catamel

LoopBack（nodejs）：企业级开发框架，由数据模型快速生成API，数据库采用MongoDB

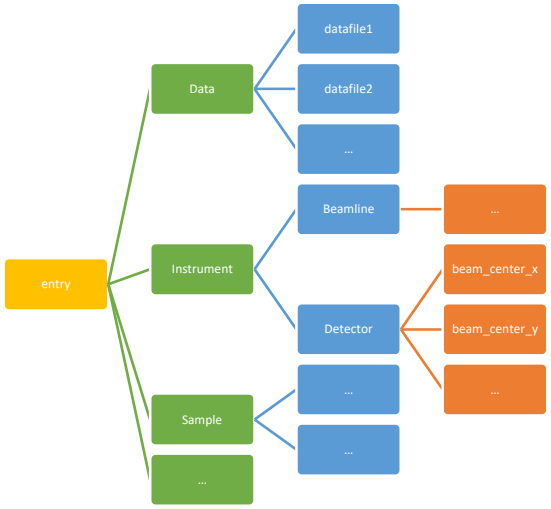
应用和开发情况：

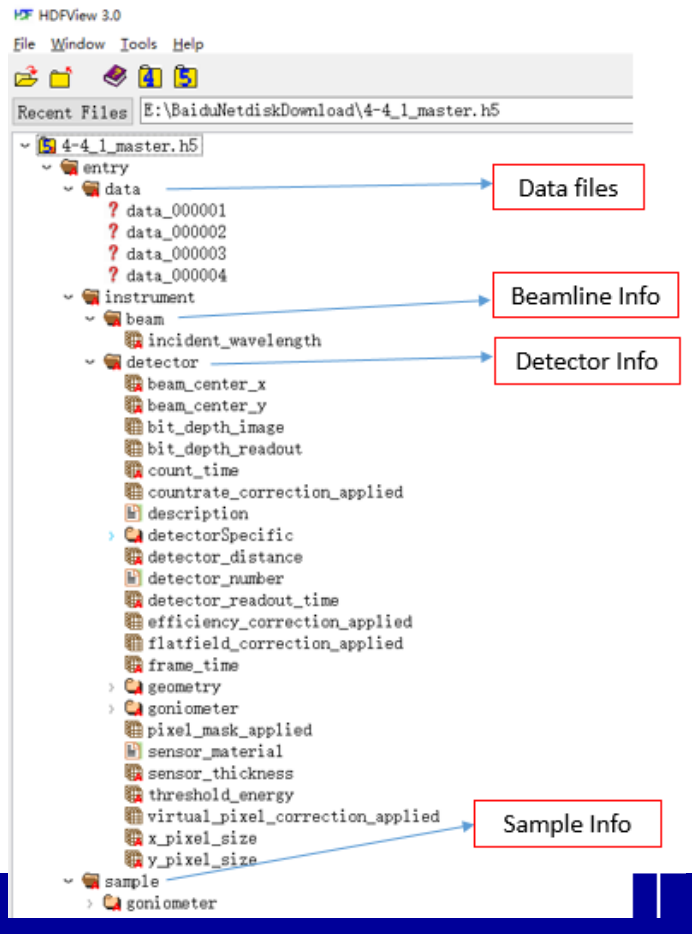
后端使用Catamel的数据模型和数据结构前端采用Java springMVC开发框架



数据格式标准化

- 数据格式现状: txt/tif/binary/dat/hdf5/…
- 利用HDF5优势： 数据和元数据一起存放
- 遵循NeXus规约—HDF5文件内容组织的规则
- HDF5 文件schema应该提前设计
- HDF5文件包含**数据+元数据**
- **Data + beamline + detector + sample+…**





3.3.1. Base Class Definitions

A description of each NeXus base class definition is given. NeXus base class d
be used in an instance of that class. Consider the base classes as a set of *con*
file.

NXaperture 孔径
A beamline aperture.

NXattenuator 衰减器
A device that reduces the intensity of a beam by attenuation.

NXbeam 线站属性
Properties of the neutron or X-ray beam at a given location.

NXbeam_stop 线站停止器
A device that blocks the beam completely, usually to protect a detector.

NXbending_magnet 弯型磁铁
A bending magnet

NXcapillary 聚焦x光线能力
A capillary lens to focus the X-ray beam.

NXcite
A literature reference

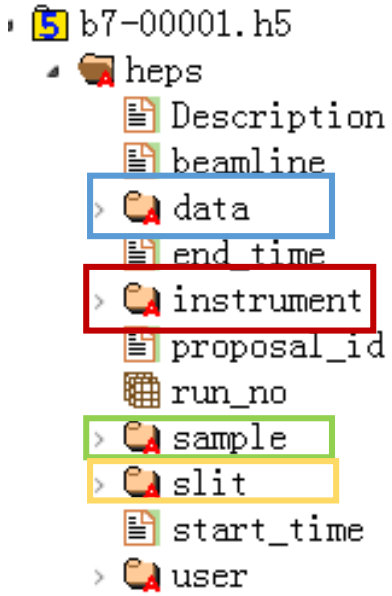
NXcollection
An unvalidated set of terms, such as the description of a beam line.

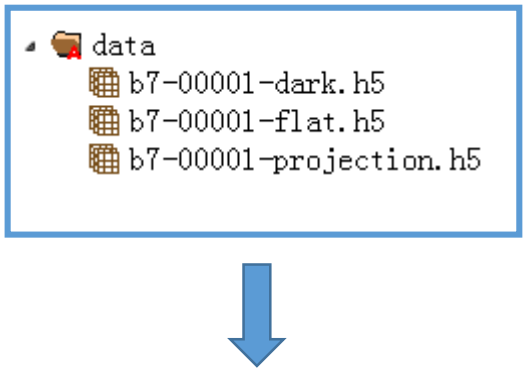
NXcollimator 准直器
A beamline collimator.

NXcrystal 晶体单色器或分析器
A crystal monochromator or analyzer.

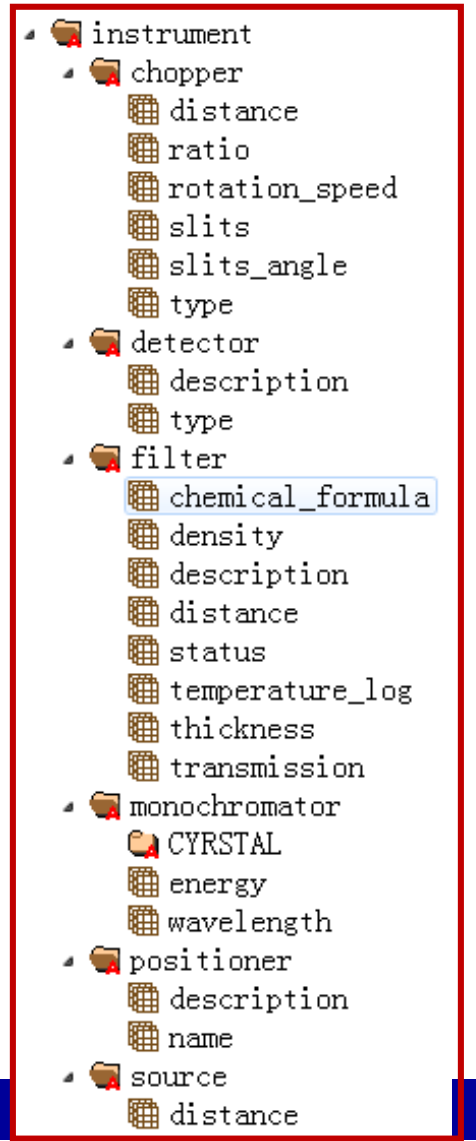
HDF5数据格式设计试点—B7成像线站

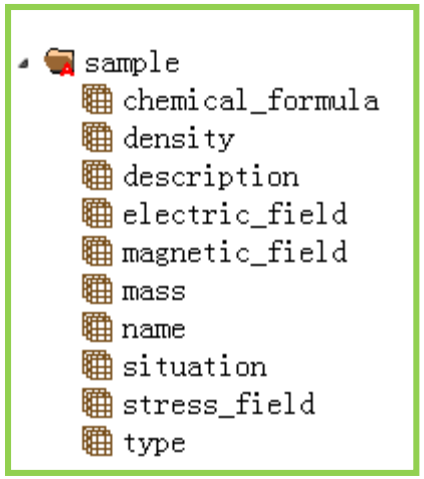
一组扫描：b7-00001.h5
b7-0001-projection.h5
b7-0001-flat.h5
b7-0001-dark.h5





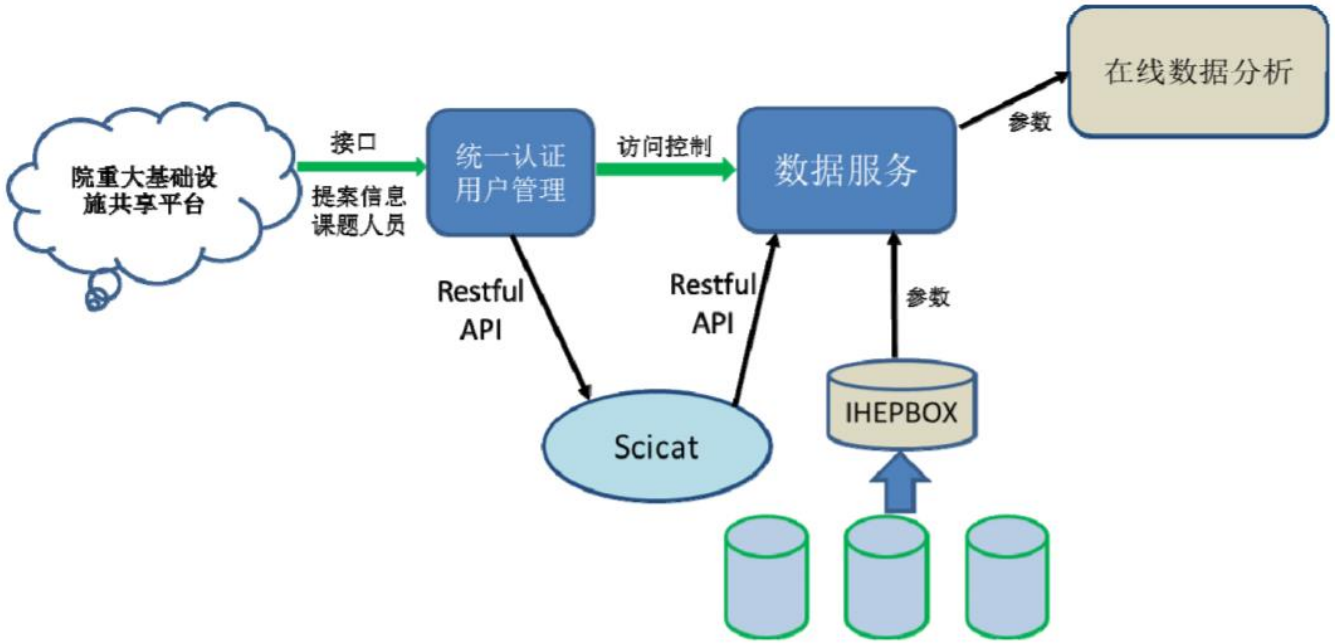
一套成像的数据，需要探测器执行三组探测。
projection：带有样品数据；
flat：不带样品数据；
dark：没有X光时的数据。

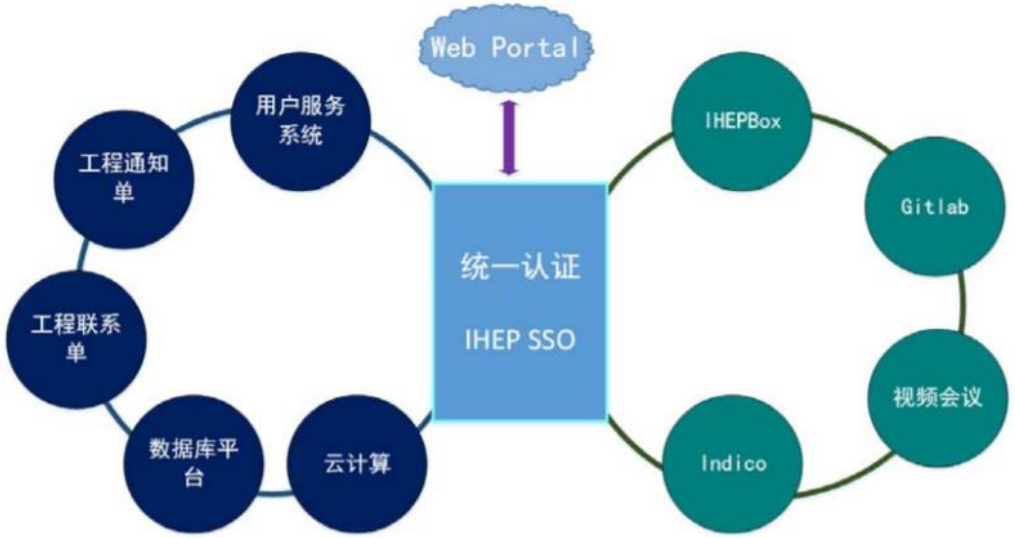




设计合理性有待讨论和验证

用户/数据服务系统





- 数据同步：院提案系统→用户服务系统
- 认证：IHEP SSO（LDAP，分组：HEPS/用户分组）
  ISSF与IHEP SSO互认证
- 实验前集群账号申请：分配gid，uid

数据服务—用户界面

Datasets / H E P S/0003 /

Details | Datafiles | Reduce | Attachments | Logbook | Lifecycle | Admin

1 datafile.

Path	Size	Time
ts/BL02/data/S1	5 GB	2019-09-20 10:50

/users/experiments/BL02/data/S1

Items per page:

Original Datablocks

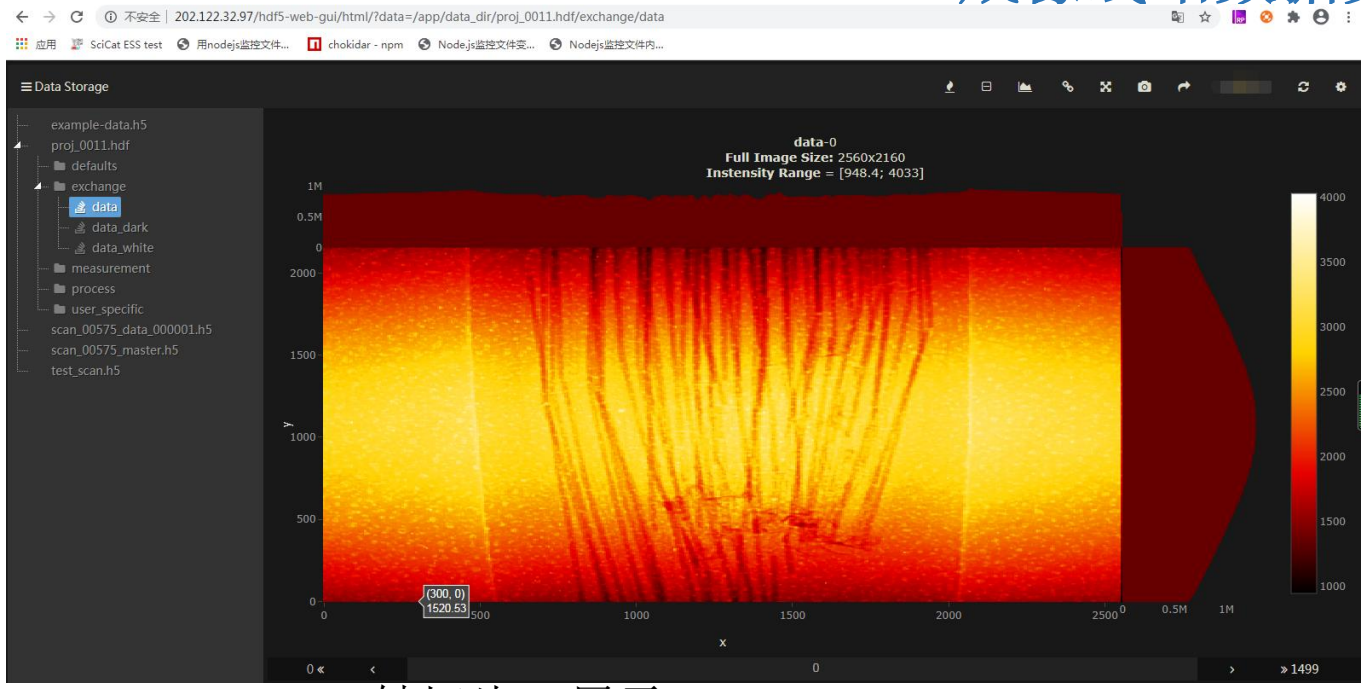
ID.	Size	Datafiles
5db14fb29ed85a4e5073ea86	9 GB	1

数据服务--数据文件web展示

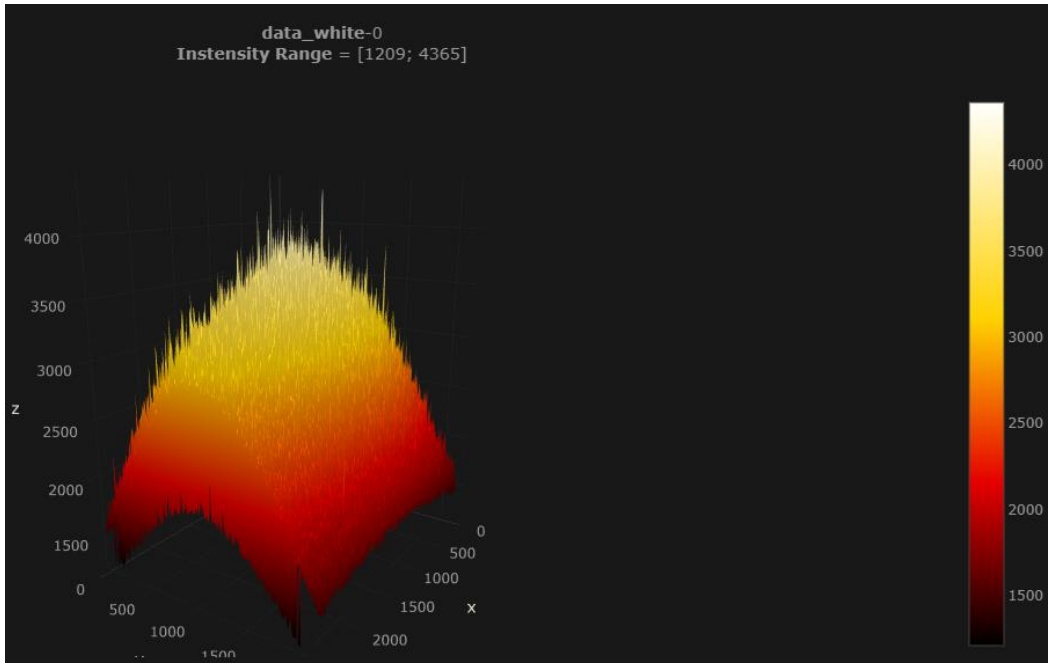
- 调研MAX_IV开发的开源的HDF API server
  - HDF5→json
  - HDF5文件树状结构
  - 图像数据展示-2D/3D
  - 缺点：无法与jupyter notebook 集成

- 调研HDFview in the cloud(**h5nuvola**)
  - front-end： HTML5, CSS, and JS
  - Back-end: Python framework Flask
  - 测试中

B7成像线站数据文件proj_0011.hdf



Z轴切片2D展示：2560*2160



3D展示：2560*2160*1500，支持拖拽

BSRF测试床建设目标

验证并优化：

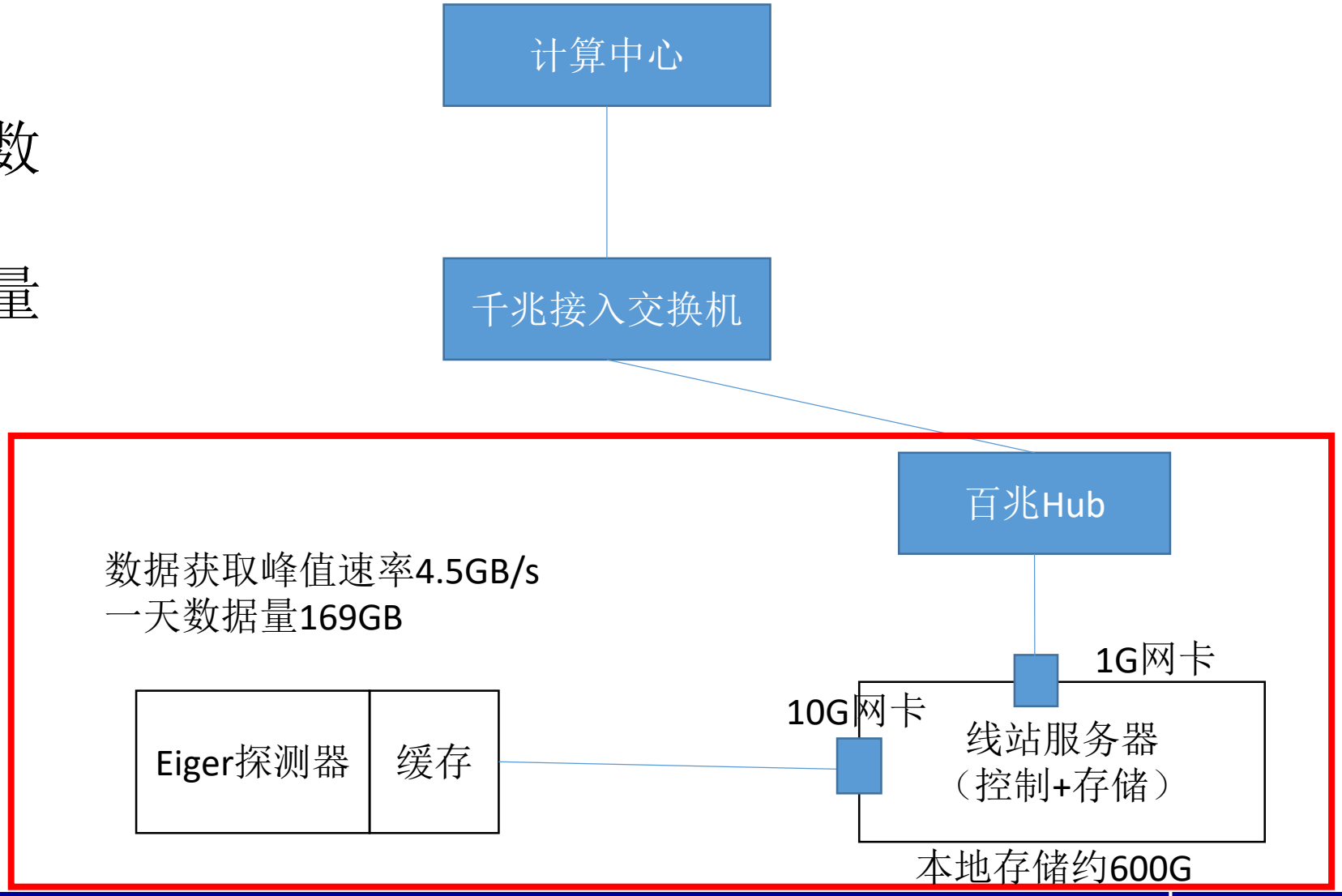
数据管理、数据传输、（线站+中心）存储、计算资源（集中管理+调度）、数据服务业务流和功能

1. 搭建满足数据采集、数据传输、数据存储、数据访问的网络环境
2. 实现采集数据的线站存储和长期集中存储
3. 实现元数据提取和数据catalogue
4. 提供基于网络的、简单的数据服务，实现数据查看、检索、下载功能
5. 实现计算资源集中管理、调度

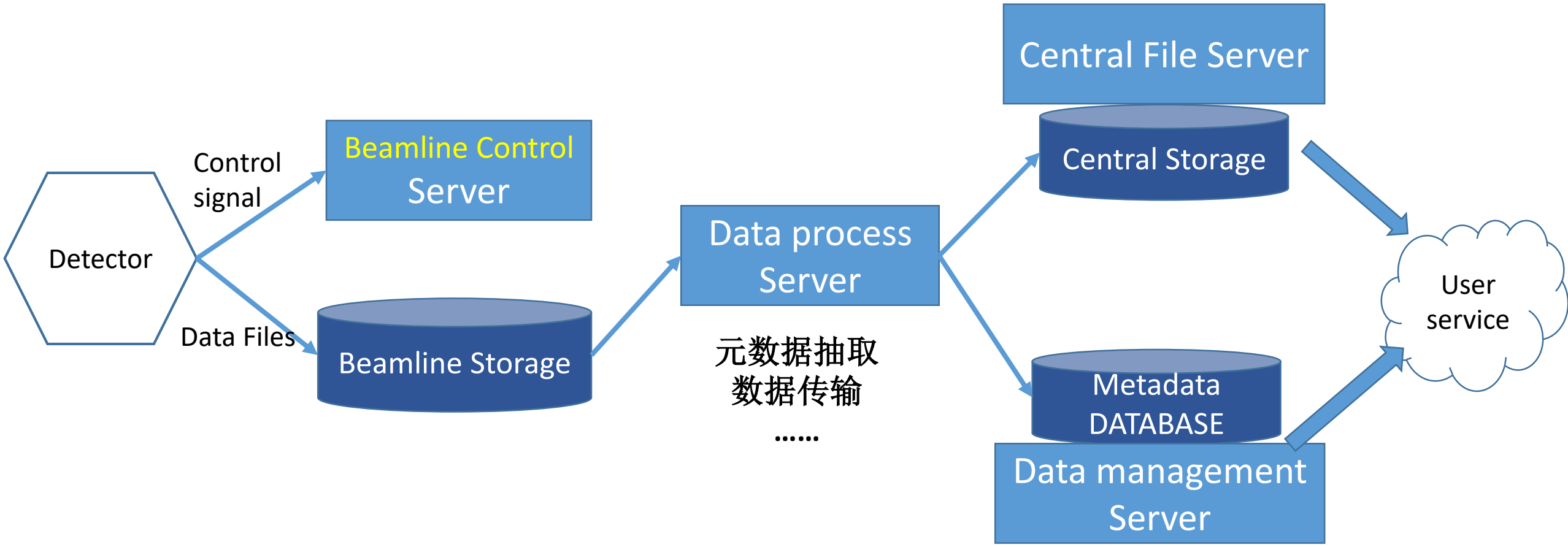
漫散射线站现状和问题

BSRF-1W1A线站问题：

1. 百兆网络环境无法满足数据传输的需求
2. 线站已有服务器磁盘容量小，无法长期保存数据



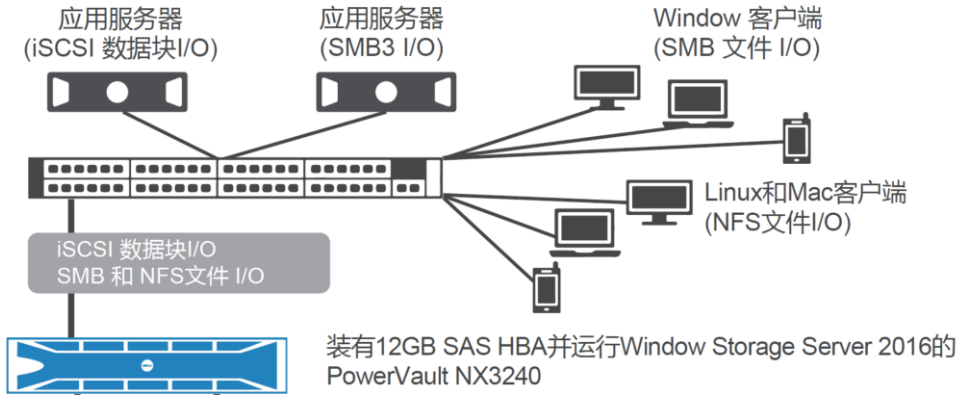
漫散射站测试床数据流


Central File Server
Central Storage
Beamline Control Server
Control signal
Detector
Data process Server
User service
Data Files
Beamline Storage
元数据抽取
数据传输
……
Metadata DATABASE
Data management Server

线站存储

- 设备选型： Dell EMC NX3240 网络存储
- 配置：
  - 两颗Intel Xeon 银牌 4214（12核处理器)
  - 64GB 内存
  - 一块以太网卡
    - 2个1 Gb Eth 口
    - 2个10Gb eth口（备用）
  - 10* 4TB NL SAS 数据盘
  - 2* 600GB SAS 系统盘
  - 支持RAID，远程监控界面

NFS不支持posix语义，影响HDF5的SWMR(单写多读)功能，
未来影响在线分析的快视功能，未来不再采用NFS
但不影响本次测试床的工作

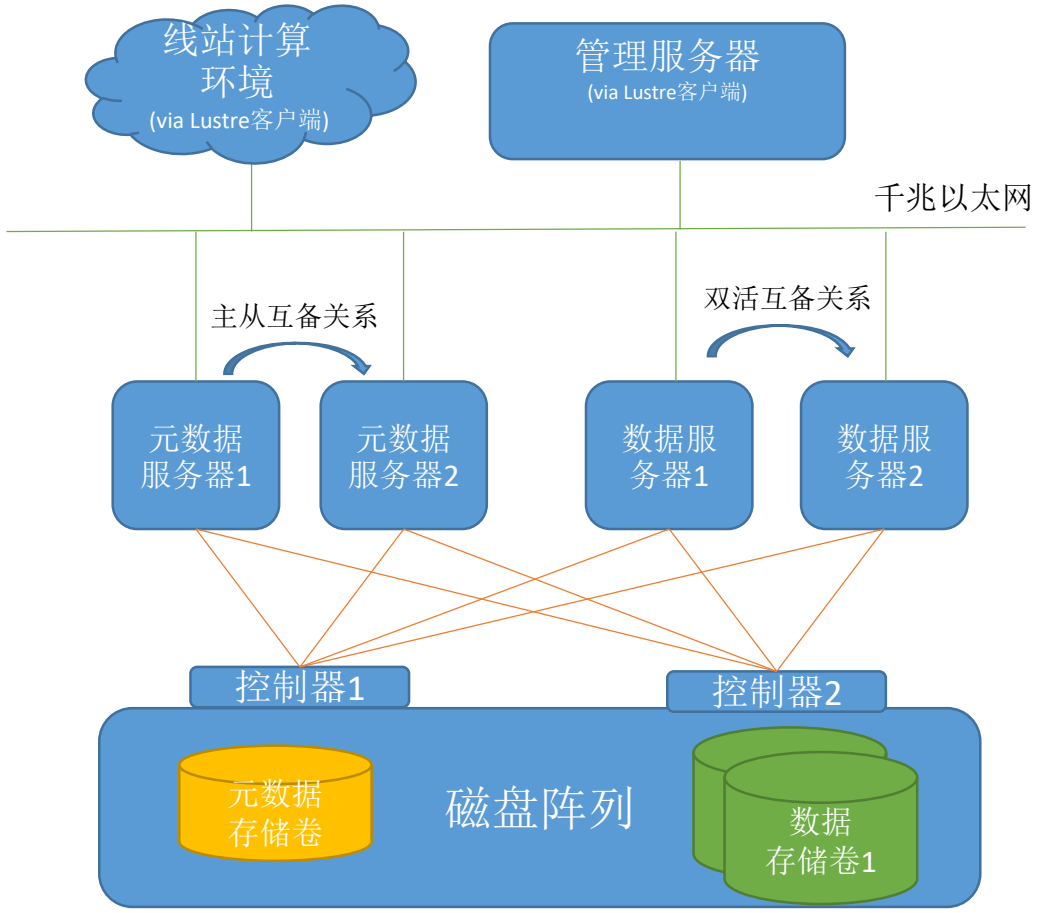


Before:
探测器数据写入线站控制服务器本地硬盘，空间有限，仅500GB，无法在网络上共享

After:
探测器数据通过线站控制服务器写入N3240，可用空间可达32TB
其它服务器可以通过NFS协议访问数据。
*峰值数据访问带宽 125MB/s，可以通过升级网络交换机，增加到 1GB/s

中心存储： 支持当前测试床，同时可以满足未来更多BSRF线站测试床数据集中存储

- 软件： 开源Lustre文件系统 2.13.x
- 4台商品服务器 （5w RMB *4 ）
  - 金牌主流Intel CPU
  - 128GB内存
  - 双口SAS数据接口卡
- 1台磁盘阵列 (20 w RMB *1)
  - 10*10TB NL SAS数据盘，可用空间80TB
  - 双控，不小于128GB控制器内存
- 存储连接：
  - 服务器通过SAS卡直连盘阵
  - 单链路物理上限(6GB/s)
- 网络连接：
  - 千兆以太网



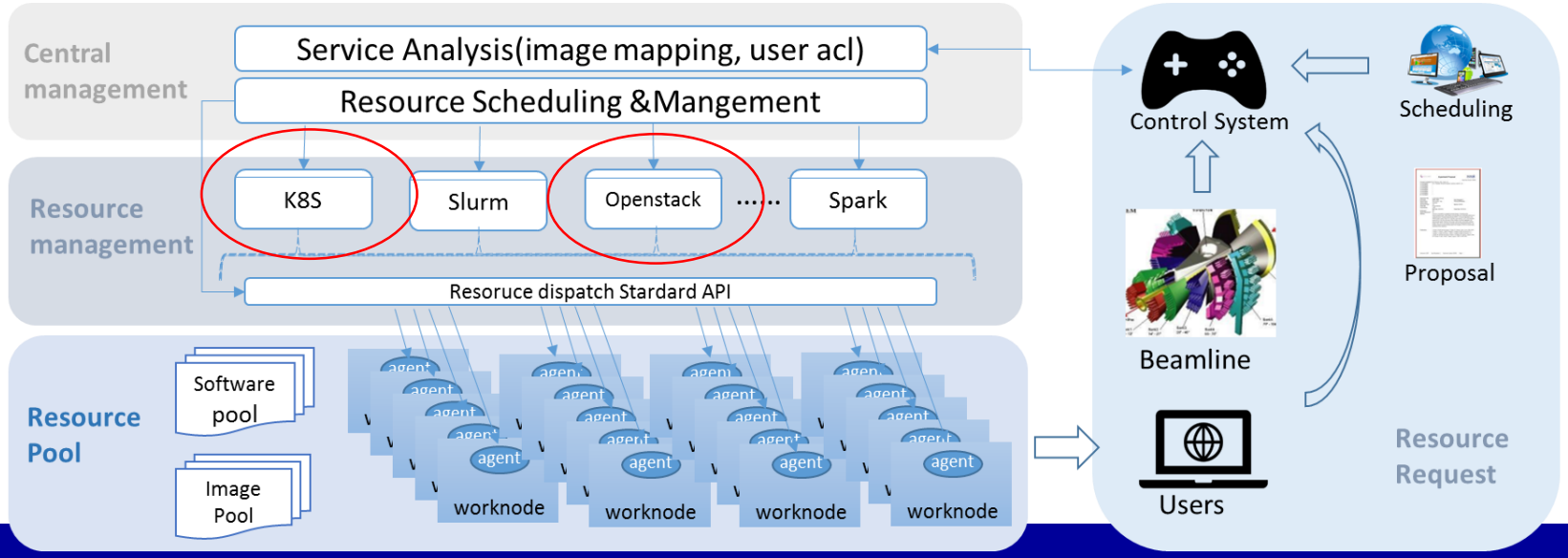
*峰值数据访问带宽 125MB/s，可以通过升级到万兆网络，增加到 1GB/s（盘阵性能限制）

计算系统：支持当前测试床，同时可以满足未来更多BSRF线站测试床计算资源需求

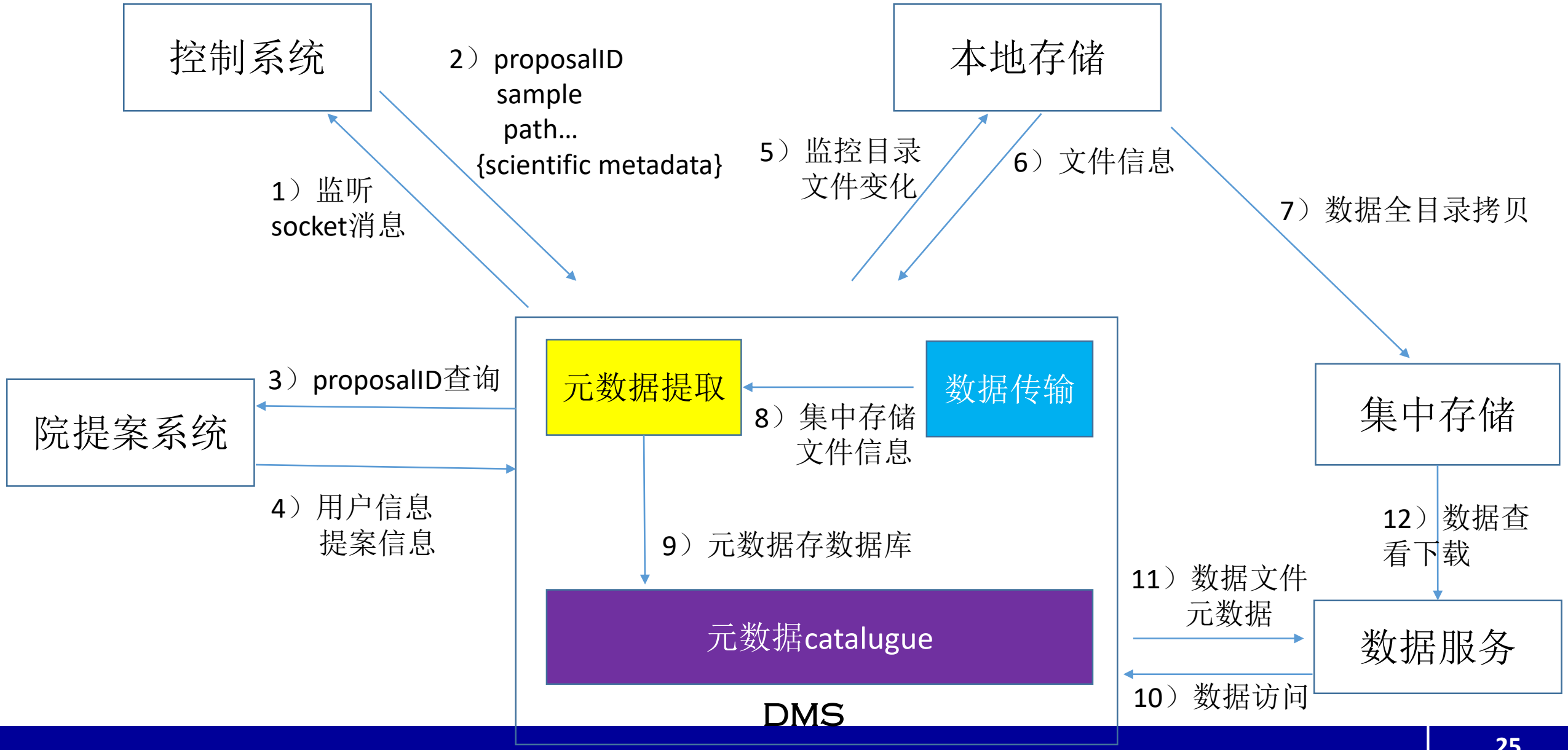
第一步：提供物理位置位于线站的在线数据处理服务器

第二步：实现计算资源的集中管理和统一调度

- 将分布在线站和计算中心的计算资源进行集中管理和统一调度
- 根据线站的需求快速透明的提供所需的计算资源
- 资源以虚拟机或容器（docker/singularity）的方式
- 虚拟机或容器集成数据处理所需的环境（操作系统类型、软件、数据等）
- 上层提供资源统一管理的界面，包括资源的物理位置，资源监控，资源使用情况等



数据管理业务流程


控制系统
本地存储
2）proposalID
sample
path…
{scientific metadata}
1）监听
socket消息
5）监控目录
文件变化
6）文件信息
7）数据全目录拷贝
元数据提取
数据传输
集中存储
3）proposalID查询
院提案系统
8）集中存储
文件信息
4）用户信息
提案信息
9）元数据存数据库
12）数据查
看下载
11）数据文件
元数据
元数据catalugue
数据服务
10）数据访问
DMS

设备清单及费用

	设备	数量	预算（万元）
网络	千兆交换机	1	0.8
存储	NAS存储	1	5.88
	开发服务器	1	3
	集中存储文件服务器	4	20
	盘阵	1	20
数据管理	数据处理服务器	1	5
		共计	54.68万元

备注：集中存储服务器和盘阵在本次testbed建设完成后，将服务于未来更多的testbed

1W1A测试床进展与问题

- 完成网络环境升级，部署线站存储和中心存储
- 完成与控制系统接口设计与研发（socket）
- 完成与传输系统接口设计与研发（mysql数据库）
- 完成元数据提取模块研发，实现元数据提取和存储
- 完成数据服务研发，实现对数据的搜索、查看和下载
- 问题：1）采用商业控制软件Spec，socket接口不可靠，发送消息前数据无法校验

  2）实验环境参数等元数据存储于文件，元数据提取工作量大

  3）线站存储不再采用NFS存储，不支持HDF5文件单写多读功能

  4）控制/DAQ最好能具备DAQ数据库，保证数据传输完整性

计划

- 3W1高数据通量线站测试床搭建
  - 采用Bluesky控制系统，设计接口
  - 将计算系统和软件框架纳入进来，验证完整业务流程
  - HDF5数据格式，元数据定义和设计
- 形成两套方案--元数据获取和catalog
  - 商业探测器控制系统
  - 自研探测器，Bluesky控制系统

谢谢！