RHIC Data and Analysis Preservation Round Table

06/26/2025

Introduction & some notes from previous meeting

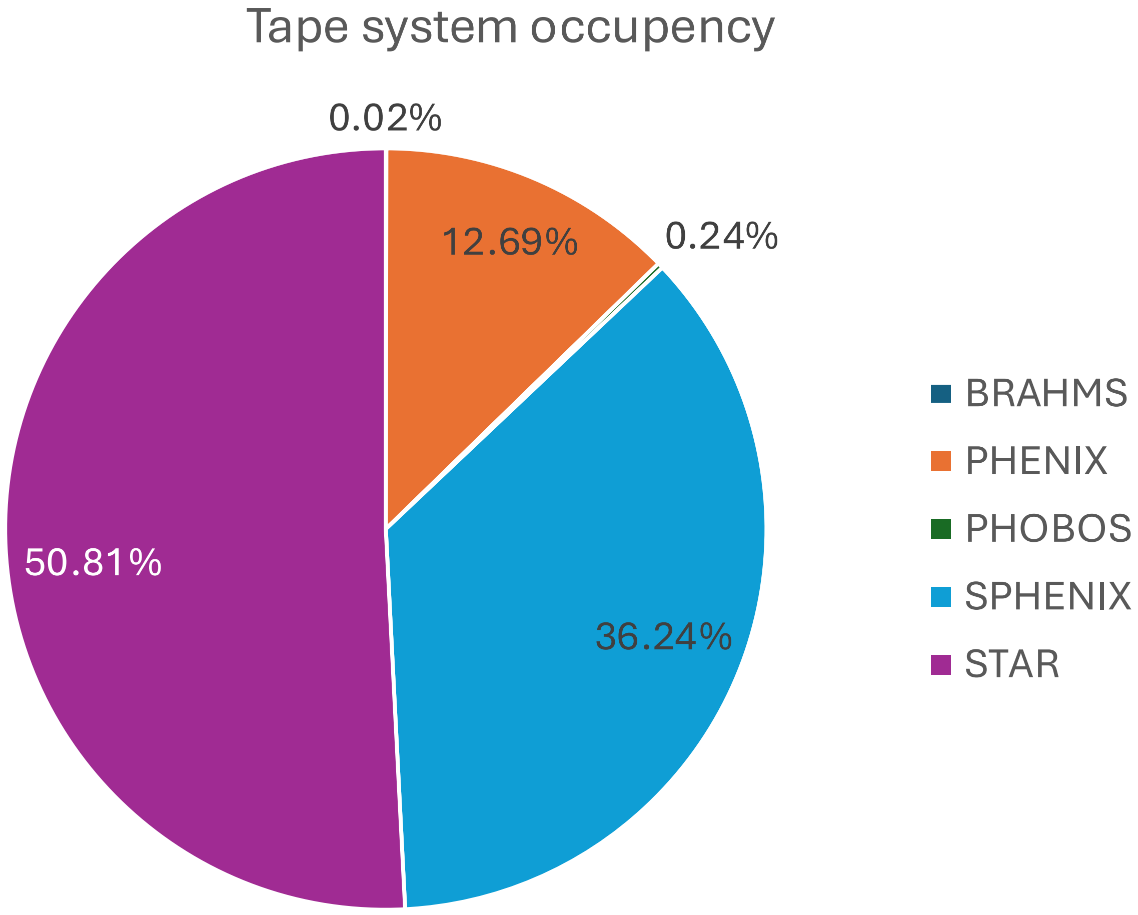
Introduction

- **Today’s focus:** review of the DAPP, document is here and linked to the indico’s page

Where are BRAHMS and PHOBOS data?

They are in HPSS

BRAMS (61 TB) and PHOBOS (604 TB) will follow the preservation path of other data on HPSS



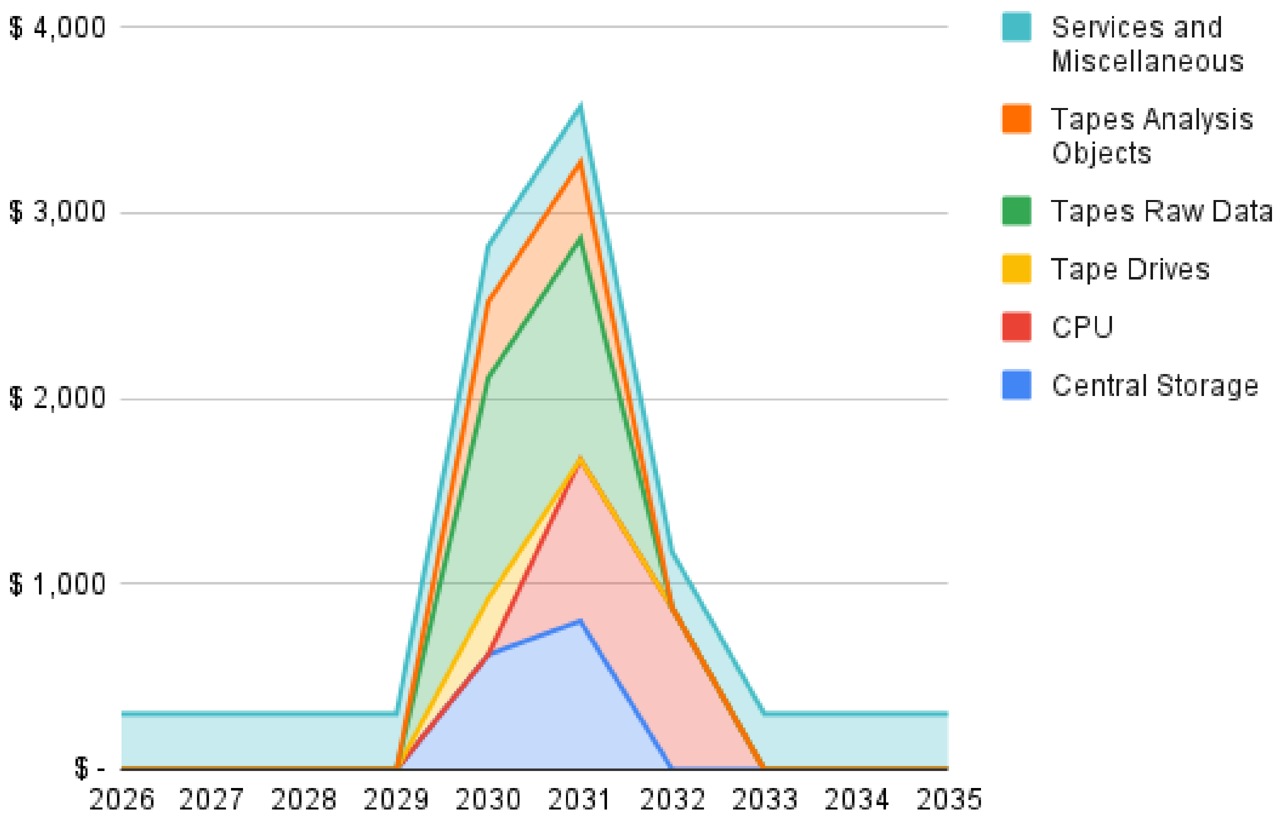
Budget profile

Budget estimates have been cross-checked.

Pie chart presented last week was missing central storage

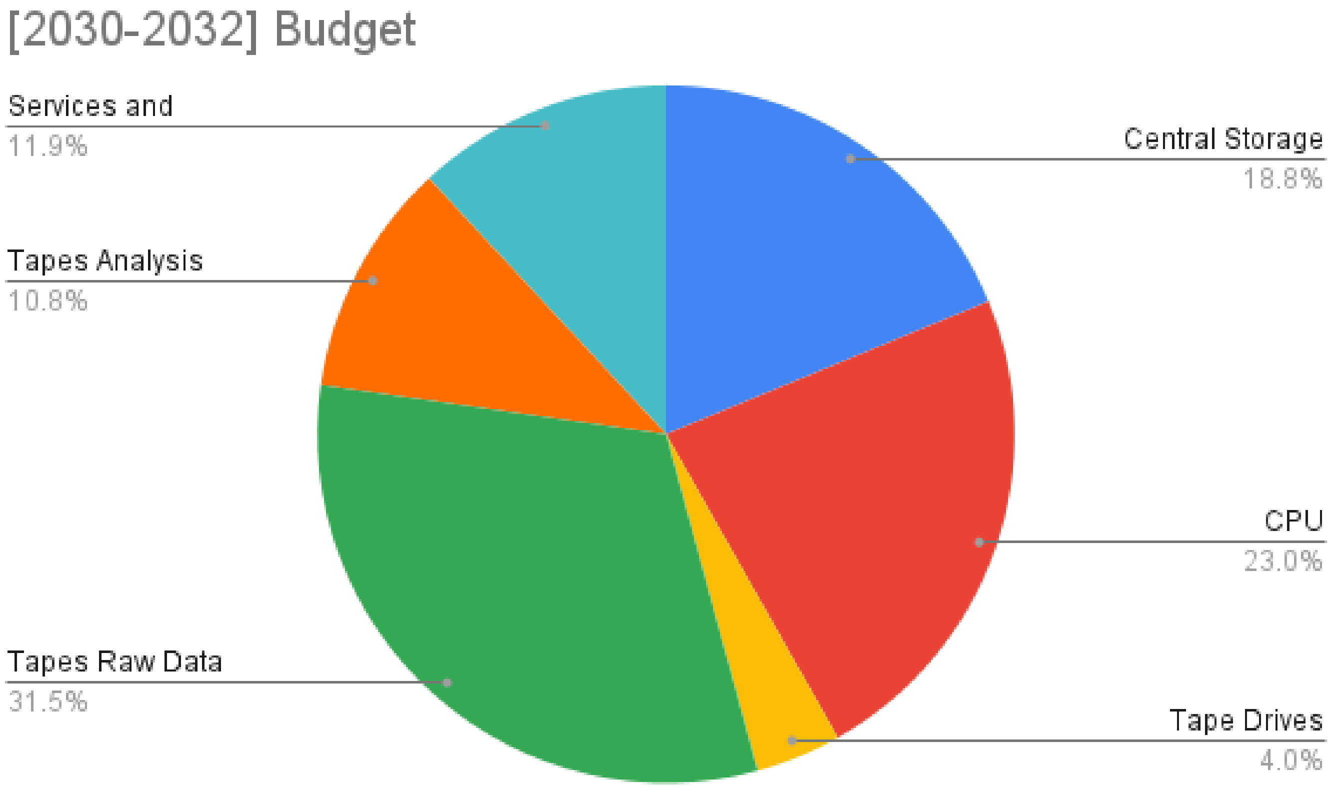
Services and Miscellaneous

- Updated since draft document
- Constant over the time period
- Provision for :
  - Cloud services (AI)
  - Allocation on CDS HPC clusters
  - Hardware for virtual services



[2030-2032] hardware budget distribution

- Corrected distribution

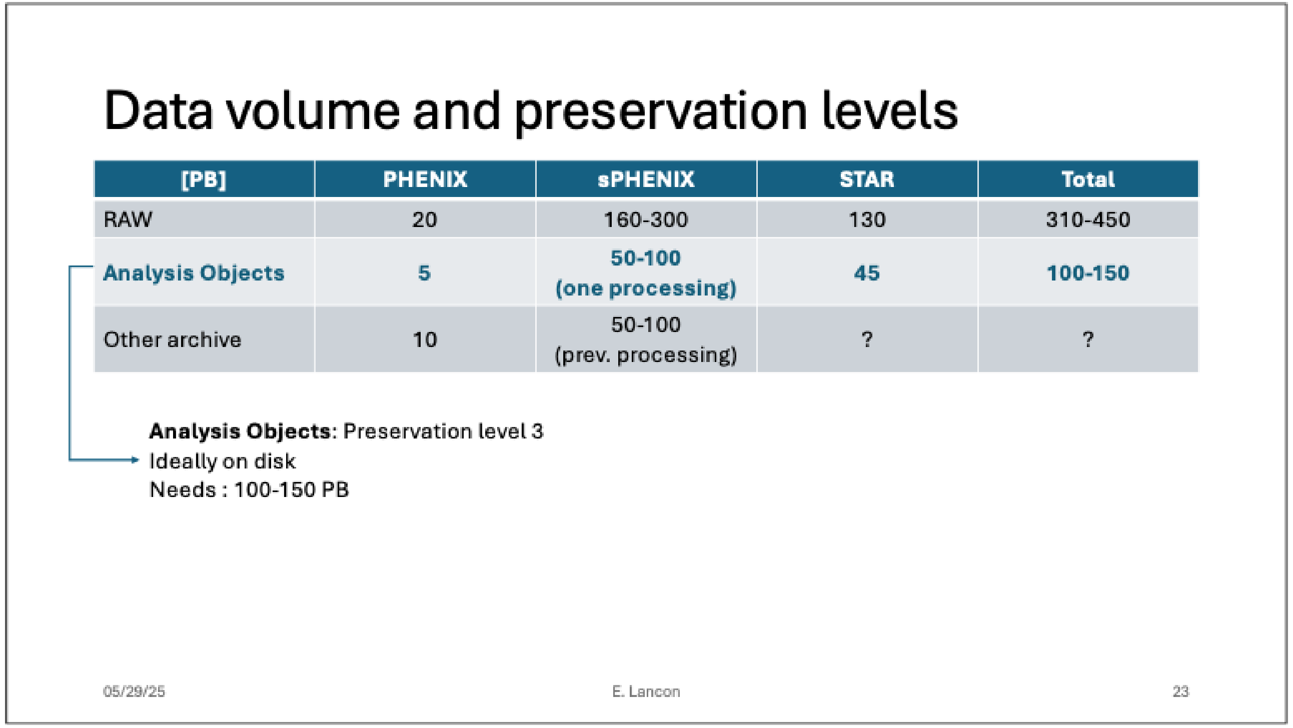


Tape systems

Preliminary estimates

- One copy of RAW and
- One copy of Analysis Objects
- RAW: 585 PB (from HPC archive + run 2025 estimate)
  Analysis Objects: 195 PB
- 45 + 15 tape drives

Slide from 05/29 meeting



Data volume and preservation levels

[PB]	PHENIX	sPHENIX	STAR	Total
RAW	20	160-300	130	310-450
Analysis Objects	**5**	**50-100** **(one processing)**	**45**	**100-150**
Other archive	10	50-100 (prev. processing)	?	?

Analysis Objects: Preservation level 3
Ideally on disk
Needs : 100-150 PB

05/29/25 E. Lancon 23

Cost of Storing 100 PB in Cloud Cold Storage

Provider	Storage Tier	$/GB-month	Monthly Cost (100 PB)	Annual Cost (100 PB)	Notes
Amazon S3	Glacier Flexible Retrieval	$0.0036	$360,000	$4,320,000	Retrieval fees apply
Amazon S3	Glacier Deep Archive	$0.00099	$99,000	$1,188,000	Cheapest archival tier; higher retrieval cost
Google Cloud	Coldline Storage	$0.004	$400,000	$4,800,000	Minimum 30 days retention
Google Cloud	Archive Storage	$0.0012	$120,000	$1,440,000	Long-term archival tier
Microsoft Azure	Cool Blob Storage	$0.0025	$250,000	$3,000,000	Minimum 30 days retention
Microsoft Azure	Archive Blob Storage	$0.00099	$99,000	$1,188,000	Lowest tier, highest latency

700 PB : 8.5 – 30 $M/y

Cost of Storing 100 PB in Cloud in Medium-Cold Storage

Provider	Storage Tier	Price (/GB/month)	Monthly Cost	Annual Cost
Amazon Web Services (AWS)	S3 Standard-Infrequent Access (S3 Standard-IA)	$0.0125	$1.25 million	$15 million
Microsoft Azure	Blob Storage – Cool Tier	$0.01	$1.00 million	$12 million
Google Cloud Platform (GCP)	Coldline Storage	$0.007	$700,000	$8.4 million

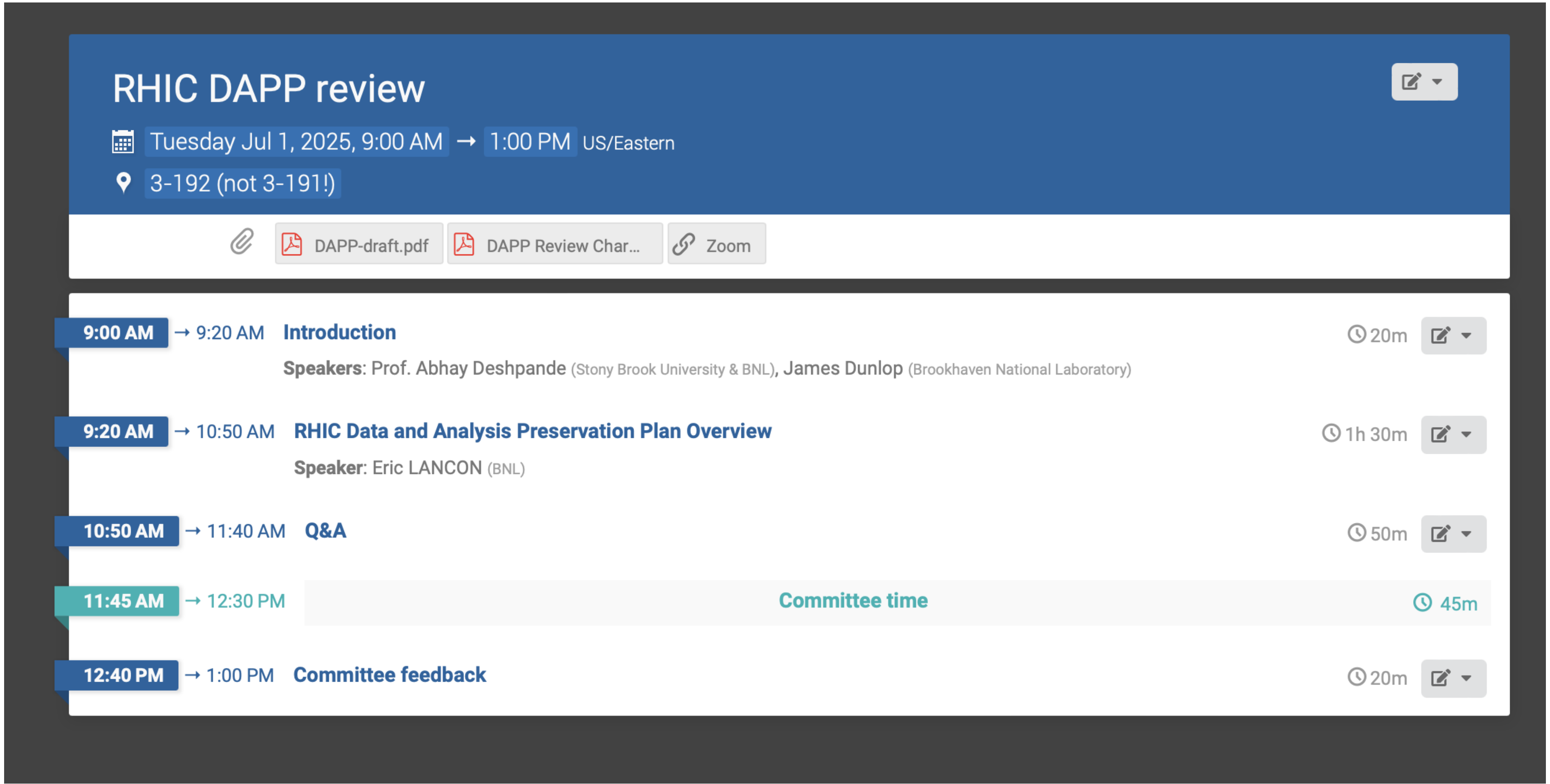
Cost of storing 100 PB in cloud tiers designed for infrequent access, a middle ground between hot and archival storage.

30 PB : 2.5 – 4.5 $M/y

Coming DAPP review

- Review Tuesday 7/1 morning (starting a 9 AM, **3-192**)
- Everyone is welcome
- It is a review of the *aspirational* Data and Analysis Preservation Plan charged by Abhay
- Indico: https://indico.bnl.gov/event/28709/

Indico’s page



RHIC DAPP review

Tuesday Jul 1, 2025, 9:00 AM → 1:00 PM US/Eastern

3-192 (not 3-191!)

DAPP-draft.pdf | DAPP Review Char... | Zoom

9:00 AM → 9:20 AM **Introduction** 20m

Speakers: Prof. Abhay Deshpande (Stony Brook University & BNL), James Dunlop (Brookhaven National Laboratory)

9:20 AM → 10:50 AM **RHIC Data and Analysis Preservation Plan Overview** 1h 30m

Speaker: Eric LANCON (BNL)

10:50 AM → 11:40 AM **Q&A** 50m

11:45 AM → 12:30 PM **Committee time** 45m

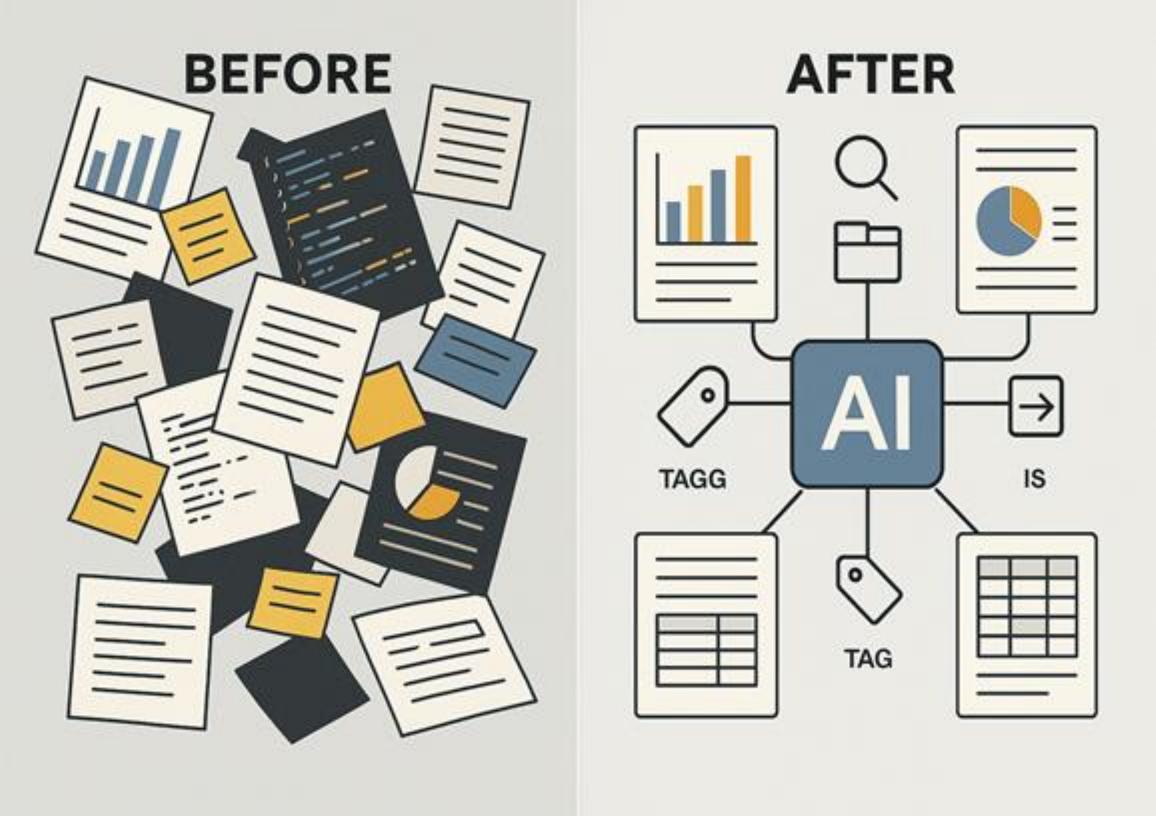
12:40 PM → 1:00 PM **Committee feedback** 20m

Slides for the review

- Once DAPP document is updated
- I’ll share draft slides for the review
- Book time tomorrow afternoon to go over the slides.

The end

AI-Powered Preservation Solutions



Objective: Lower the barrier to RHIC data reuse by assisting users—especially non-experts and newcomers—in finding and understanding data and workflows.

Why AI?

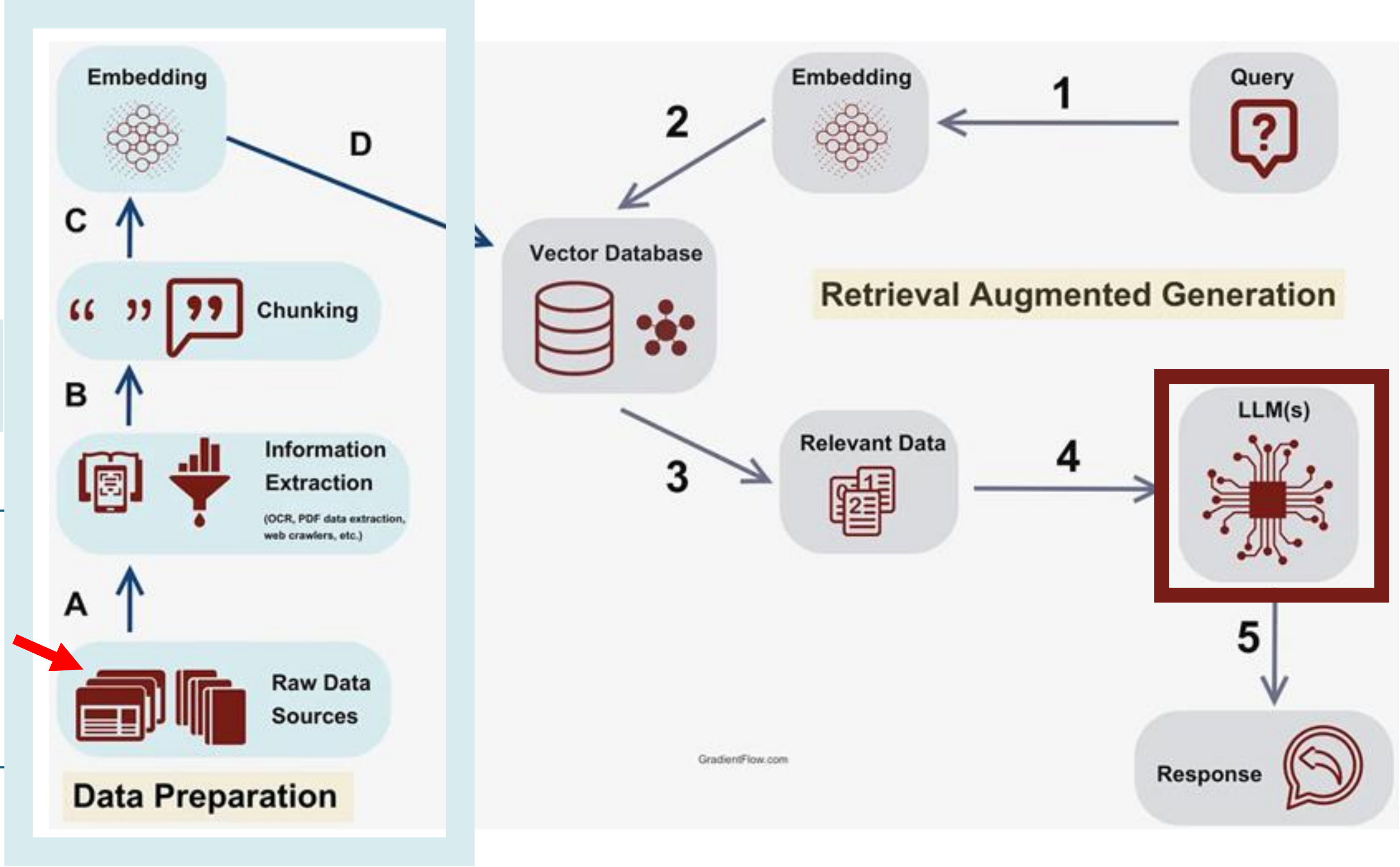
- **AI Capabilities**
  - Semantic search across experiments' documentation and records (both public and internal).
  - Retrieval-Augmented Generation (RAG) for accurate, contextual answers
  - Assistance in finding relevant datasets, papers, and workflows
  - Suggestions for code snippets or container workflows
- Finding the dataset is one thing. Knowing how it was processed, with which software version and parameters, is another. AI helps connect these dots automatically.
- Natural language assistant based on RHIC content, the ChatBot
- **Benefits**
  - Speeds up learning curve for new users
  - Allows educators to extract curated teaching materials

Retrieval Augmented Generation (RAG)

Domain-specific knowledge

Published data

Cleaned and validated internal data



LLM designed for general use

https://gradientflow.substack.com/p/best-practices-in-retrieval-augmented

Recursive Multi-Format Web-Content Extraction Framework

- **Multi-format scraping**
  - Parses HTML, PDF, Word (DOC/DOCX), PowerPoint (PPT/PPTX), and EPS/PS files in one pass with on-the-fly document conversion
  - Cleanses formats and boilerplate yielding high-quality plain text
  - Captures metadata (URL, page title, and timestamp) and text into yaml
  - Extensible

- **Recursive crawl with smart filters**
  - Traverses a site depth-first while skipping unwanted paths and file types
  - Dynamic filtering keeps it inside the seed domain.

- **Respectful rate-limiting**: configurable SLEEP delay prevents overwhelming servers.

- **Extension-aware analytics and report**:
  - Tracks processed, skipped, redirected, and dynamically-filtered link
  - Reports grouped stats for quick checks and decision-making

www.star.bnl.gov/

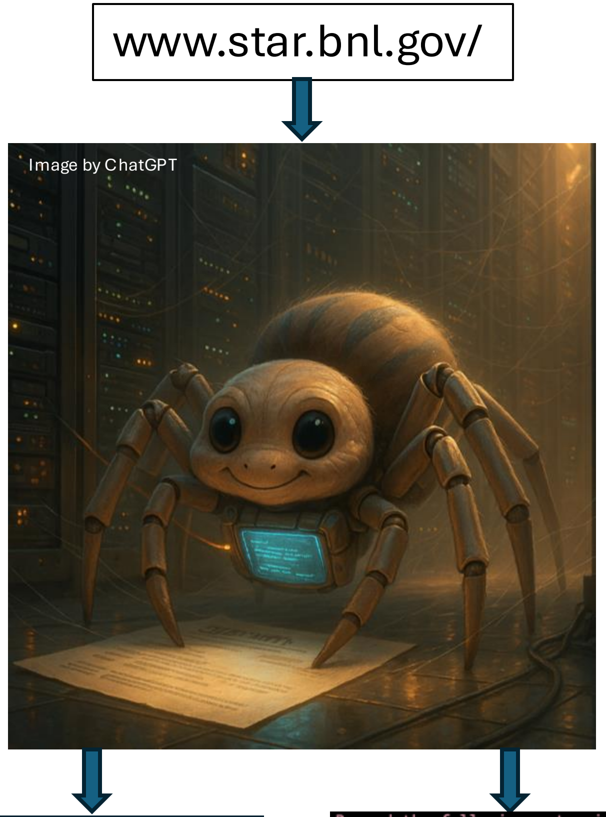
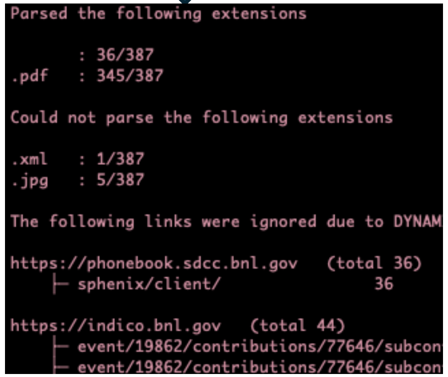


Image by ChatGPT

DataBase



```
Parsed the following extensions

        : 36/387
.pdf    : 345/387

Could not parse the following extensions

.xml    : 1/387
.jpg    : 5/387

The following links were ignored due to DYNAM

https://phonebook.sdcc.bnl.gov   (total 36)
     ├─ sphenix/client/                  36

https://indico.bnl.gov   (total 44)
     ├─ event/19862/contributions/77646/subcon
     ├─ event/19862/contributions/77646/subcon
```

Review of DAPP

Chair:

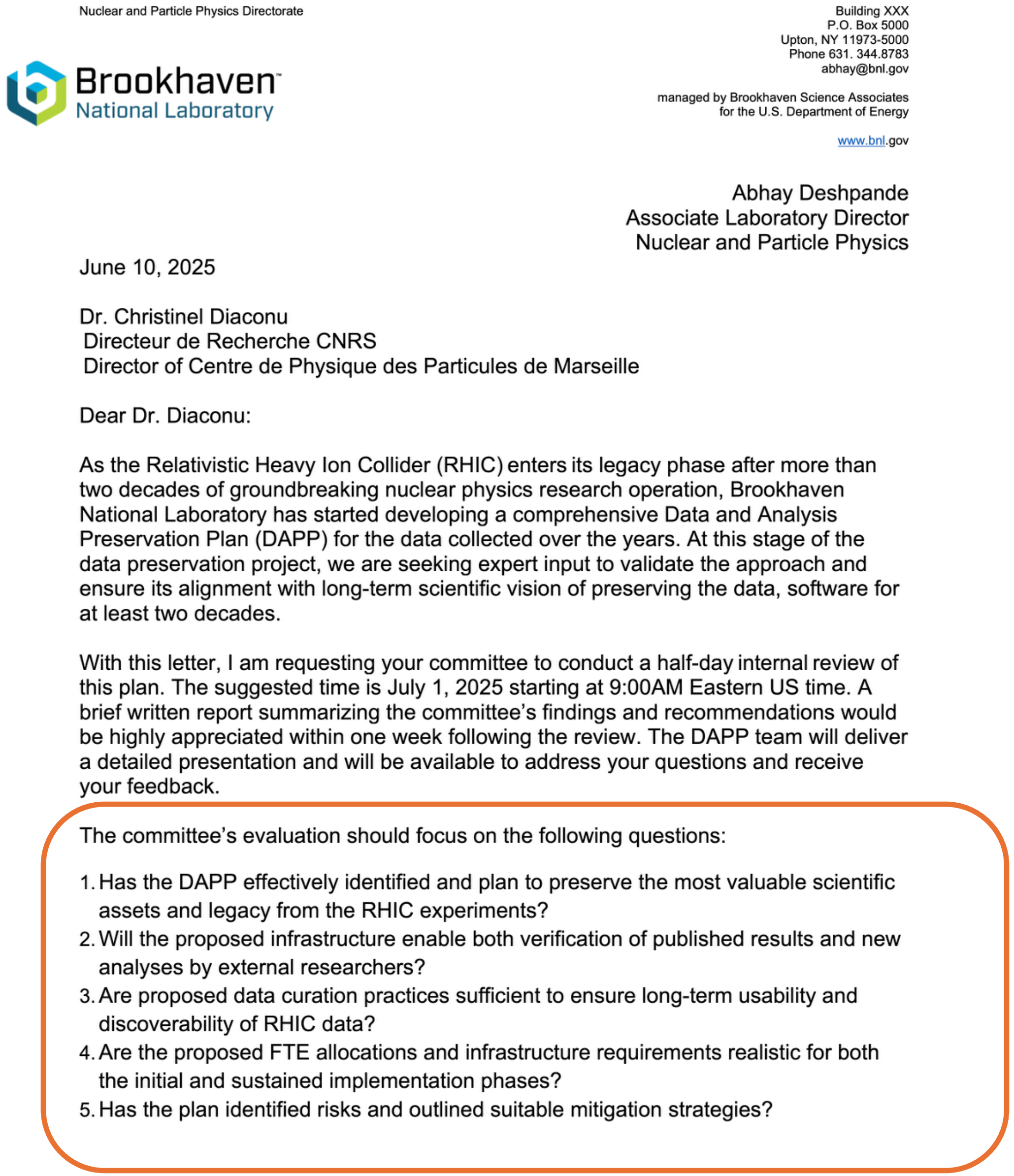
Cristel Diaconu – H1 / DPHEP Chair

Members

- Simone Campana – CERN
- Achim Geiser – ZEUS / CMS
- Kati Lassila-Perini – ICFA Data Life Cycle
- Ulrich Schwickerath – CERN Preservation
- Ralf Seidl – BNL / RIKEN

Date

- **July 1st**
- Half-day (morning session)
- **Over Zoom**



Nuclear and Particle Physics Directorate

Brookhaven National Laboratory

Building XXX
P.O. Box 5000
Upton, NY 11973-5000
Phone 631. 344.8783
abhay@bnl.gov

managed by Brookhaven Science Associates
for the U.S. Department of Energy

www.bnl.gov

Abhay Deshpande
Associate Laboratory Director
Nuclear and Particle Physics

June 10, 2025

Dr. Christinel Diaconu
Directeur de Recherche CNRS
Director of Centre de Physique des Particules de Marseille

Dear Dr. Diaconu:

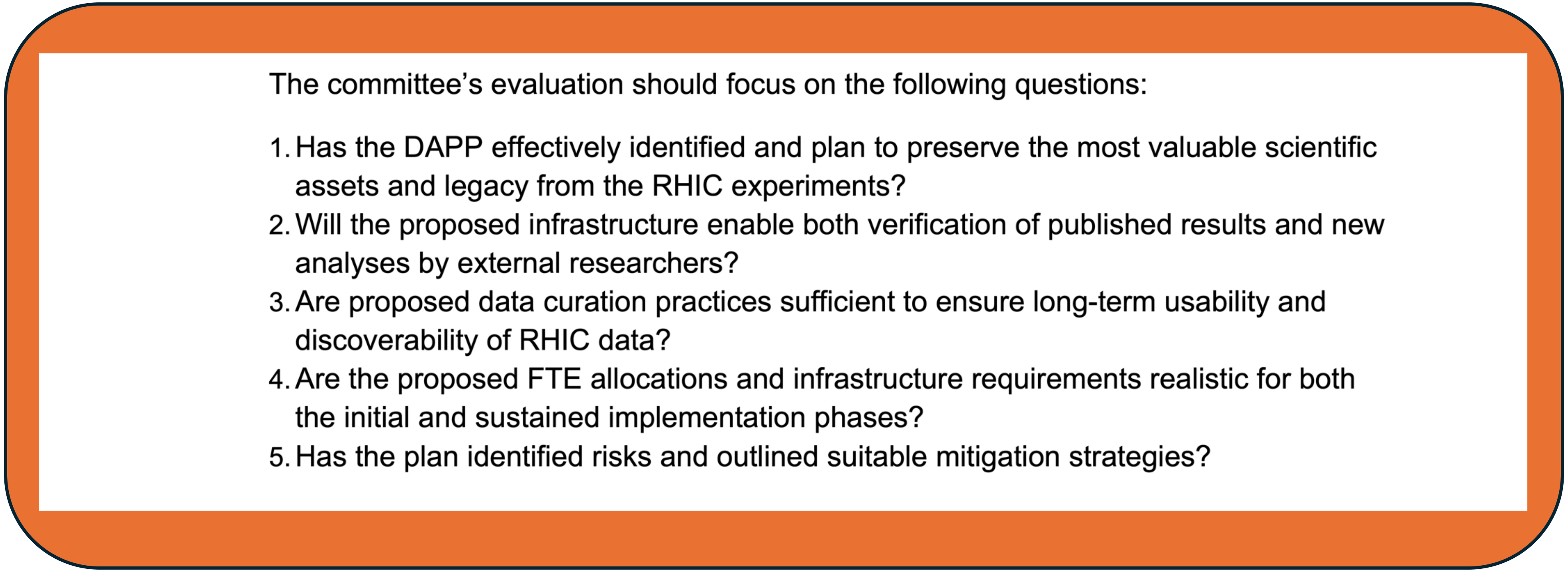
As the Relativistic Heavy Ion Collider (RHIC) enters its legacy phase after more than two decades of groundbreaking nuclear physics research operation, Brookhaven National Laboratory has started developing a comprehensive Data and Analysis Preservation Plan (DAPP) for the data collected over the years. At this stage of the data preservation project, we are seeking expert input to validate the approach and ensure its alignment with long-term scientific vision of preserving the data, software for at least two decades.

With this letter, I am requesting your committee to conduct a half-day internal review of this plan. The suggested time is July 1, 2025 starting at 9:00AM Eastern US time. A brief written report summarizing the committee's findings and recommendations would be highly appreciated within one week following the review. The DAPP team will deliver a detailed presentation and will be available to address your questions and receive your feedback.

The committee's evaluation should focus on the following questions:

1. Has the DAPP effectively identified and plan to preserve the most valuable scientific assets and legacy from the RHIC experiments?
2. Will the proposed infrastructure enable both verification of published results and new analyses by external researchers?
3. Are proposed data curation practices sufficient to ensure long-term usability and discoverability of RHIC data?
4. Are the proposed FTE allocations and infrastructure requirements realistic for both the initial and sustained implementation phases?
5. Has the plan identified risks and outlined suitable mitigation strategies?

Charge questions



The committee’s evaluation should focus on the following questions:

1. Has the DAPP effectively identified and plan to preserve the most valuable scientific assets and legacy from the RHIC experiments?
2. Will the proposed infrastructure enable both verification of published results and new analyses by external researchers?
3. Are proposed data curation practices sufficient to ensure long-term usability and discoverability of RHIC data?
4. Are the proposed FTE allocations and infrastructure requirements realistic for both the initial and sustained implementation phases?
5. Has the plan identified risks and outlined suitable mitigation strategies?