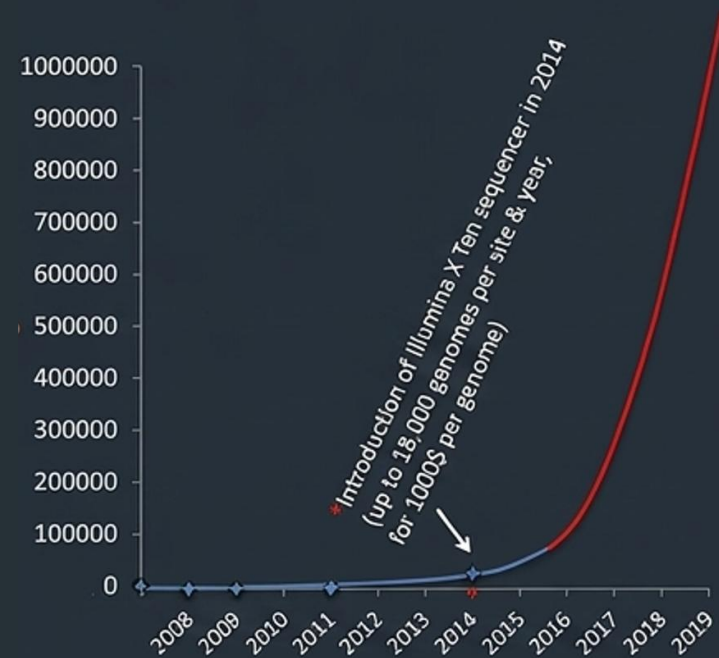


Data Platform & Engines for AI

Frank Lee PhD
Distinguished Engineer, IBM Corp
frankli@us.ibm.com | @drfranklee

Scott Pearson, Camelia Williams, Daryl Williams,

From 4V to 4A



Volume

Velocity

Variety

Veracity

Data Abstraction

Multi-cluster parallel filesystem

Multi-site data grid

Multi-cloud global name space

Data Acceleration

GPU-Direct

S3-filesystem vertical cache

Lakehouse caching

Client-side acceleration
KV Cache Offloading

Data Access

HDFS

Object/S3

CNSA/CSI

S3-Vector
API/MCP

Data Awareness

Catalog

Context

Content

Composition

2005

2010

2015

2020

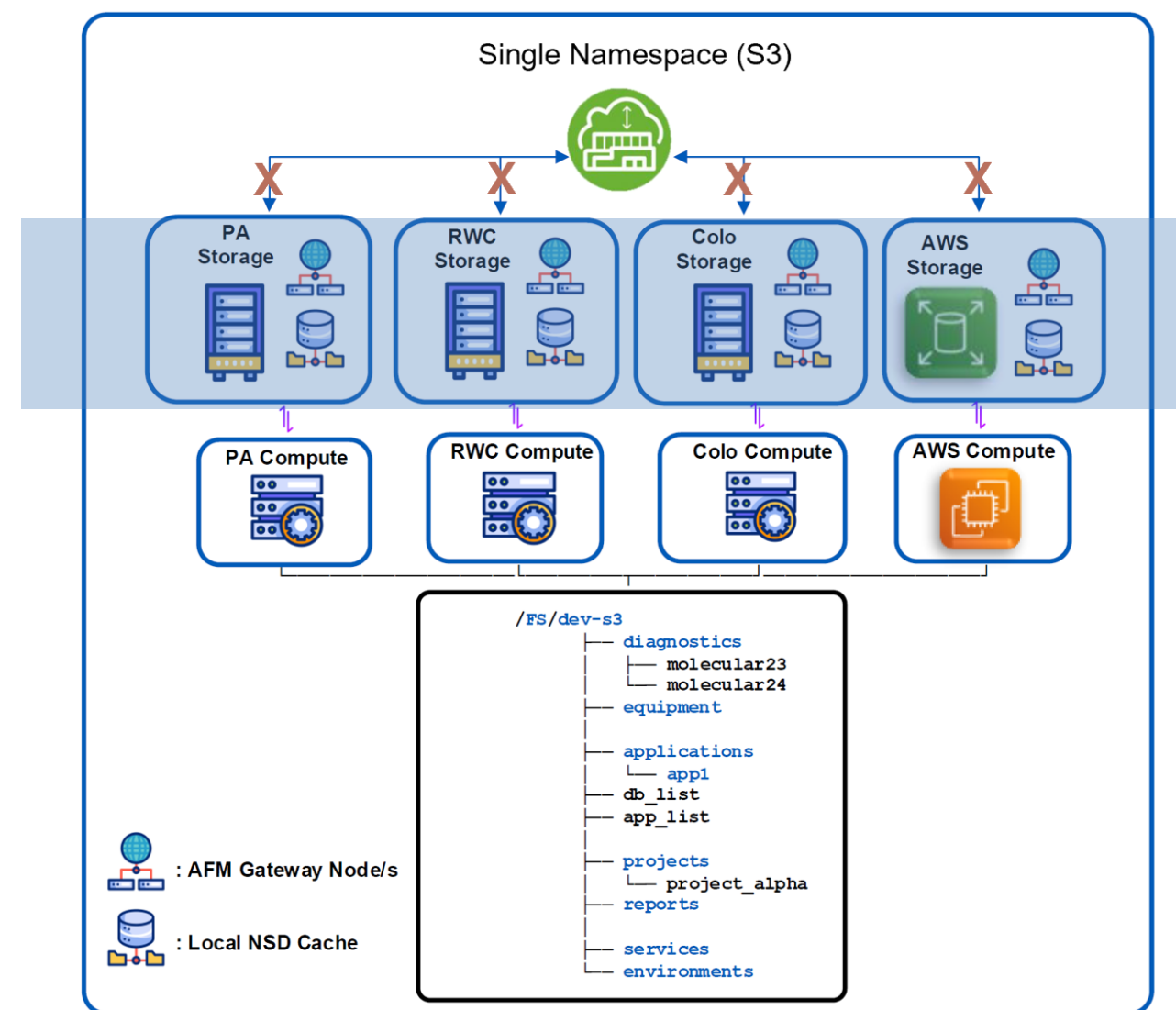
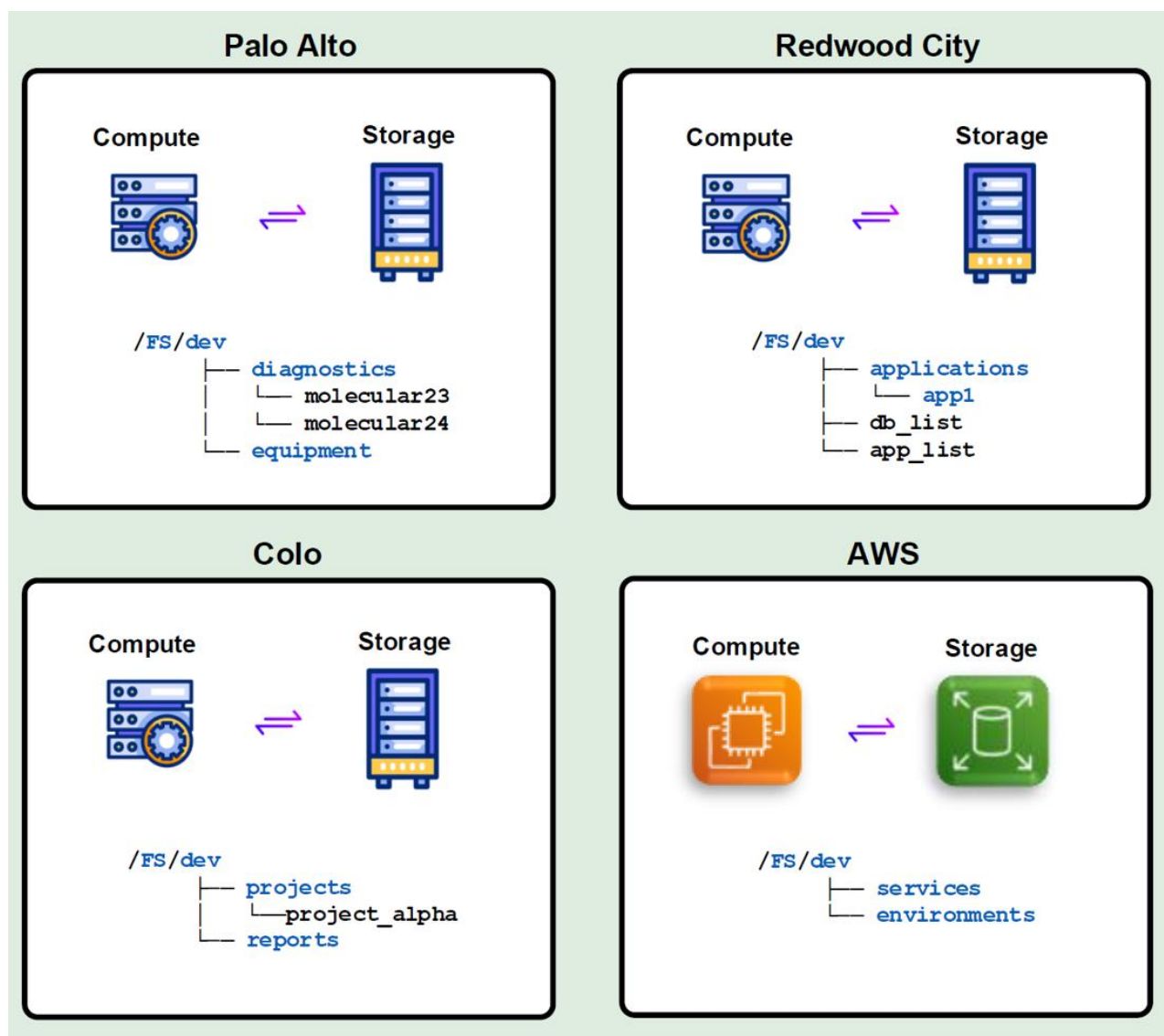
2025

HPC

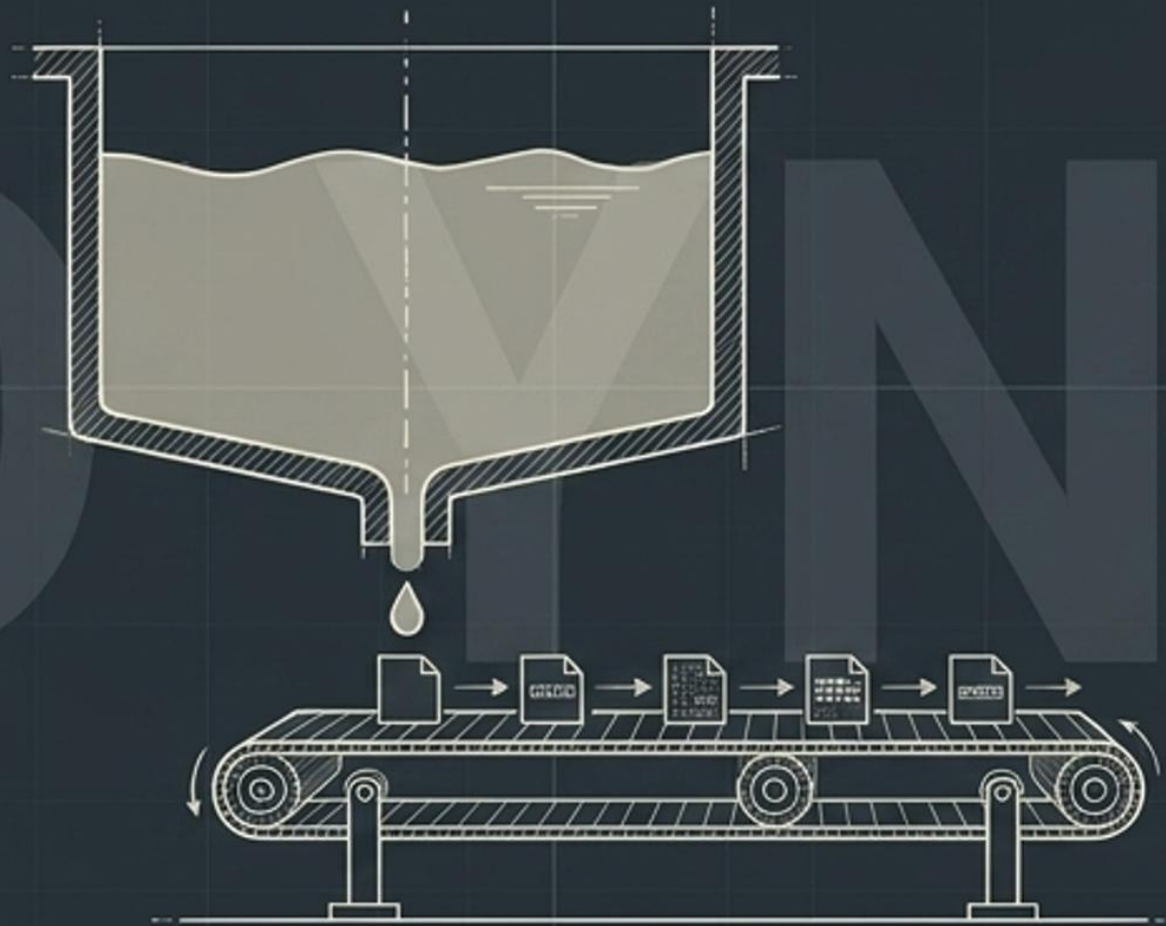
Cloud

AI

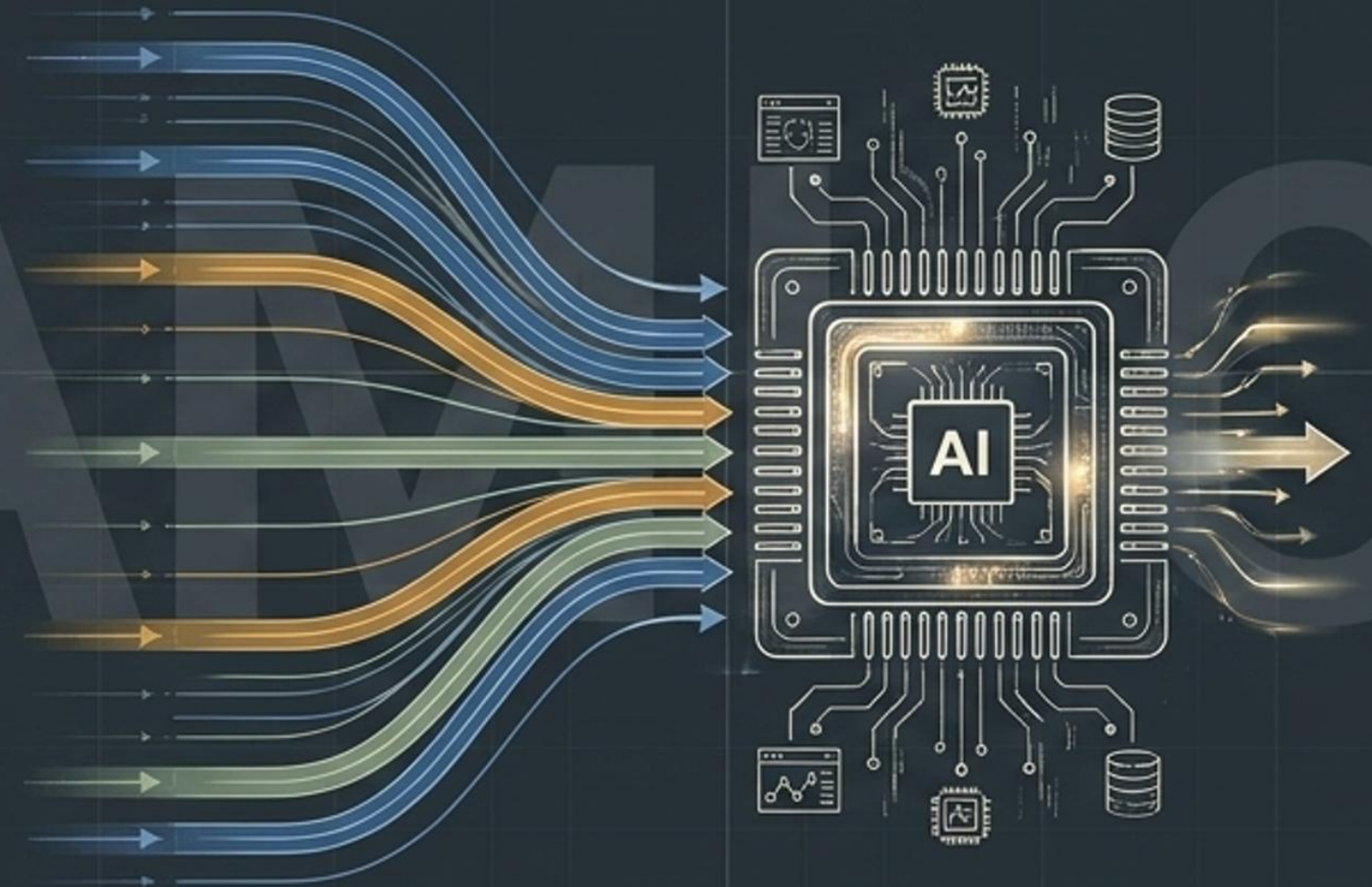
A global namespace binding all data locations



Data Acceleration (from Glacier to KV Cache)



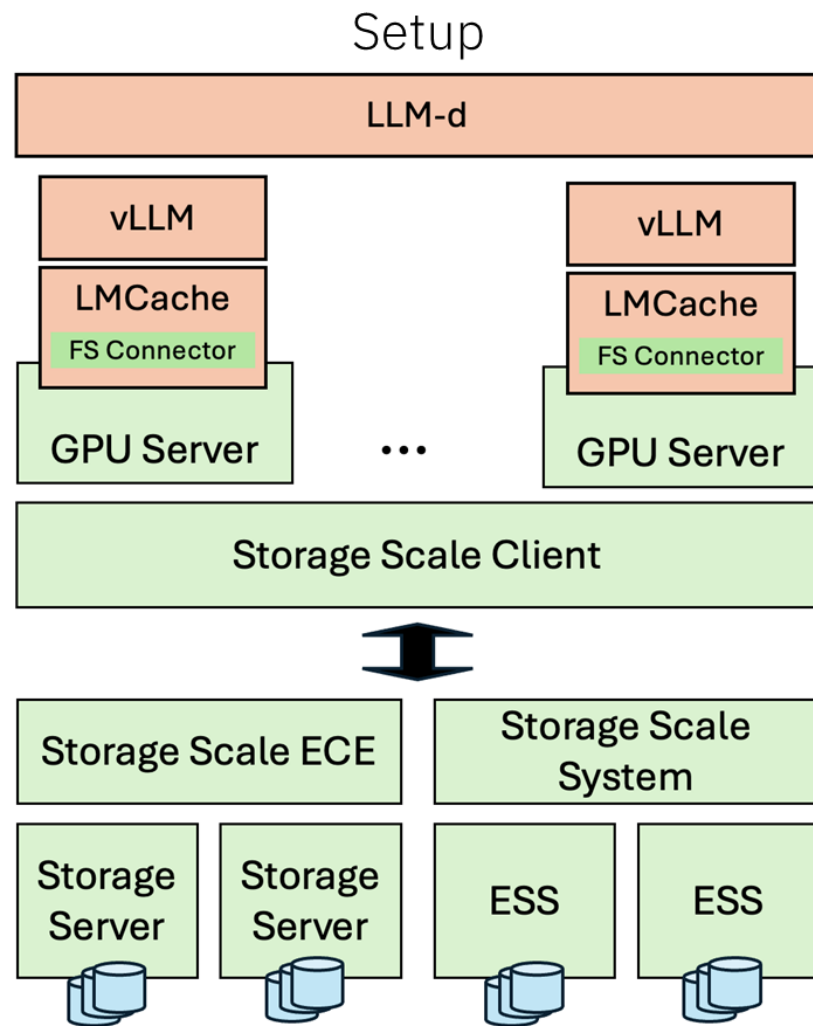
Workflow and AI (training, inferencing) driving active data usage – no more “dormant” data



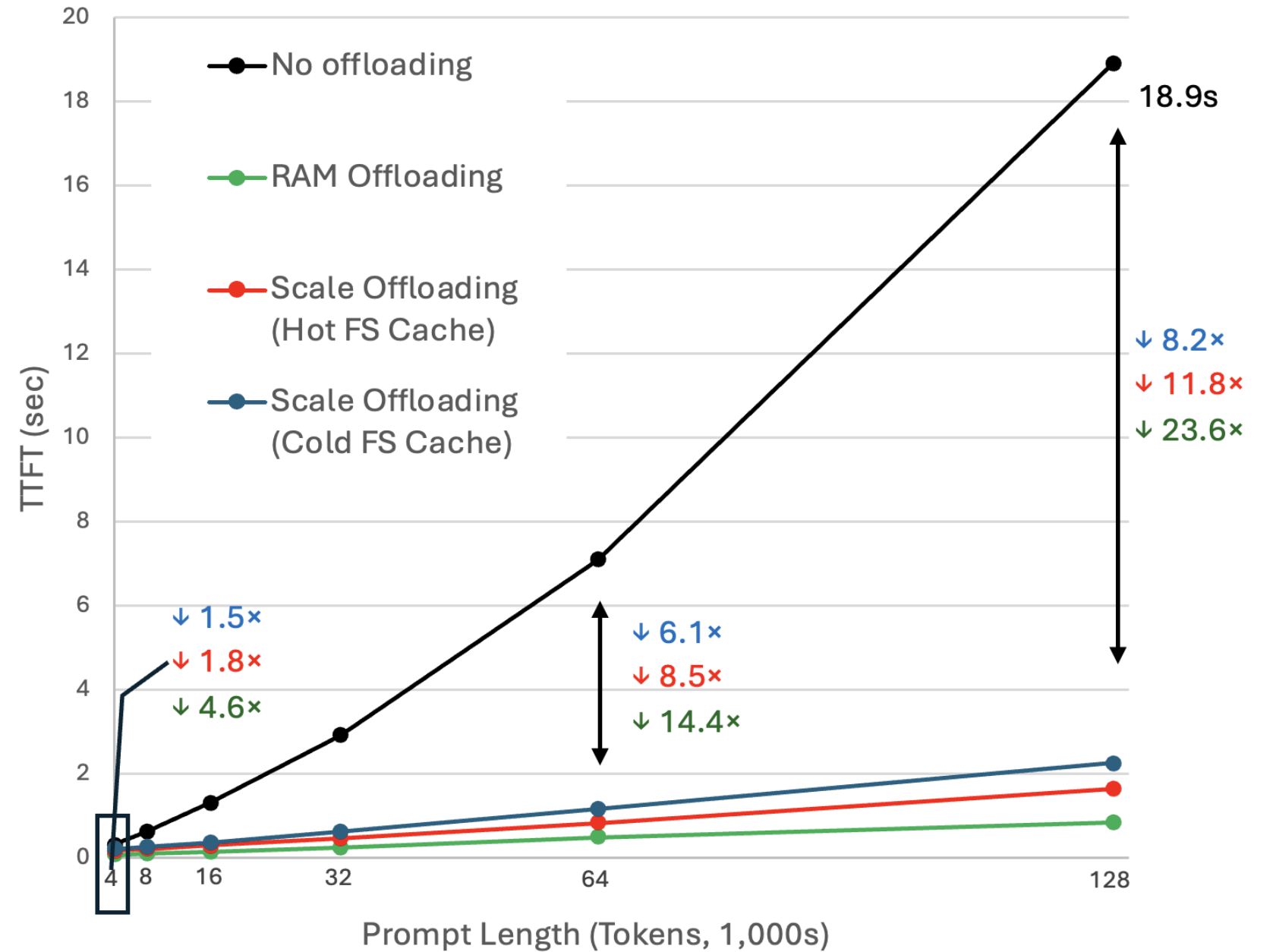
Global Data Platform accelerating data from archival storage into GPU short-term and long-term memory for biomedical research (HPC & AI Inferencing)

AI Acceleration: 10X Gain

128K AI Prompt Experiment



10X Gain in Acceleration



NOAAtempX v1.0

NOAA Global Surface Summary of the Day (GSOD) Data Dashboard

[Compare Cities](#) [Historical Trends](#)

Year

2023

Location

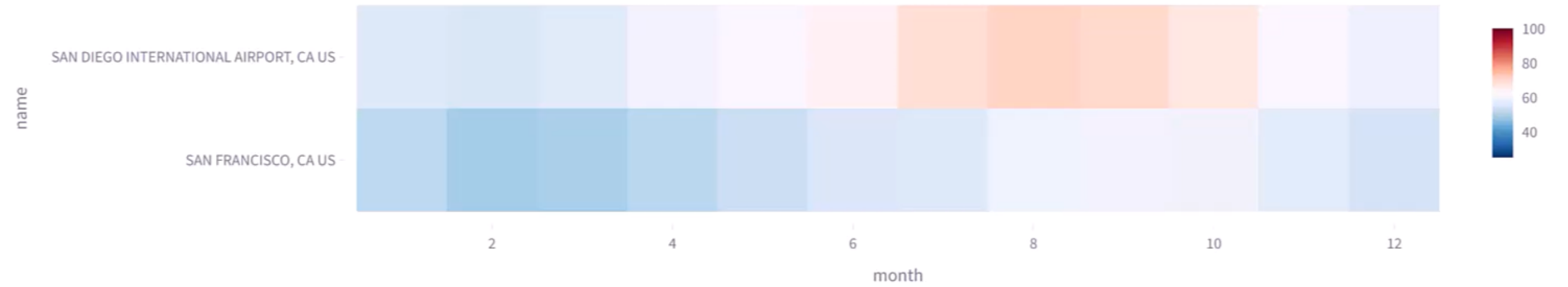
San ... x

San ... x

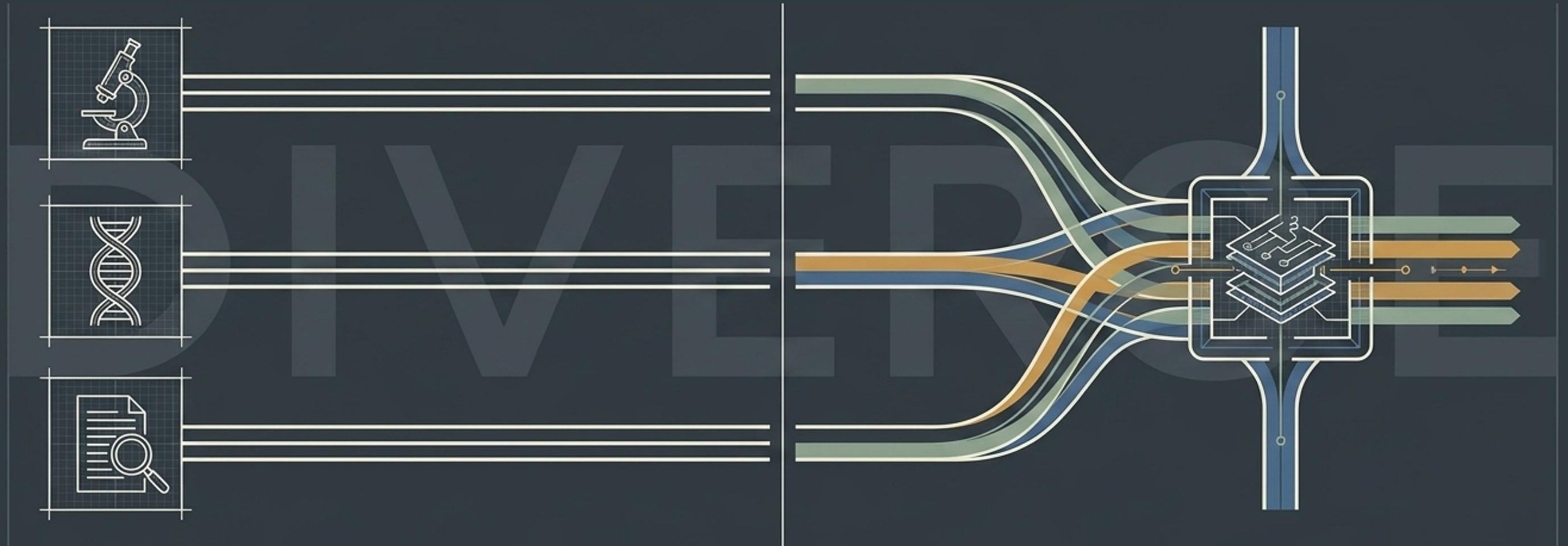
Height of Graph

150

show data table

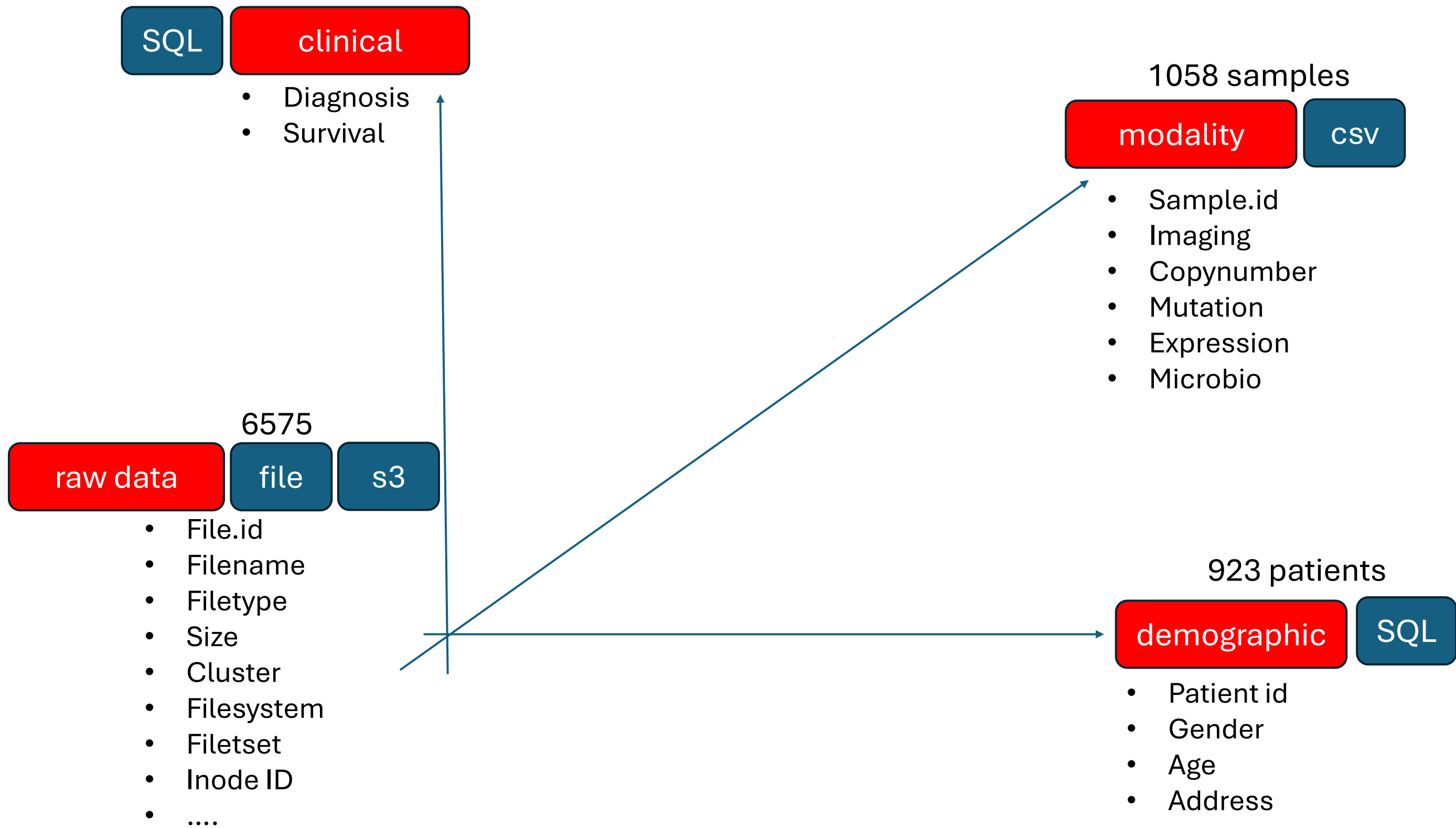


Data Access : Multi-everything (modal, protocol, endpoint)



Data/set are increasingly multi-modal, multi-type (unstructured, structured and streaming) and multi-format (file, object, vector)

Global Data Platform with data-native accessibility – multi-modal, multi-protocol (file, object, SQL), multi-endpoint (API, MCP)



Multimodal Biomedical Research Data Dashboard

Home **Visualize** Explore Search Catalog

Main-Category

- Diagnosis
- Gender
- Age
- Address
- Patient
- Sample
- Collection
- Full-col-
count

Cross-Category

Gender



Height of Graph

540

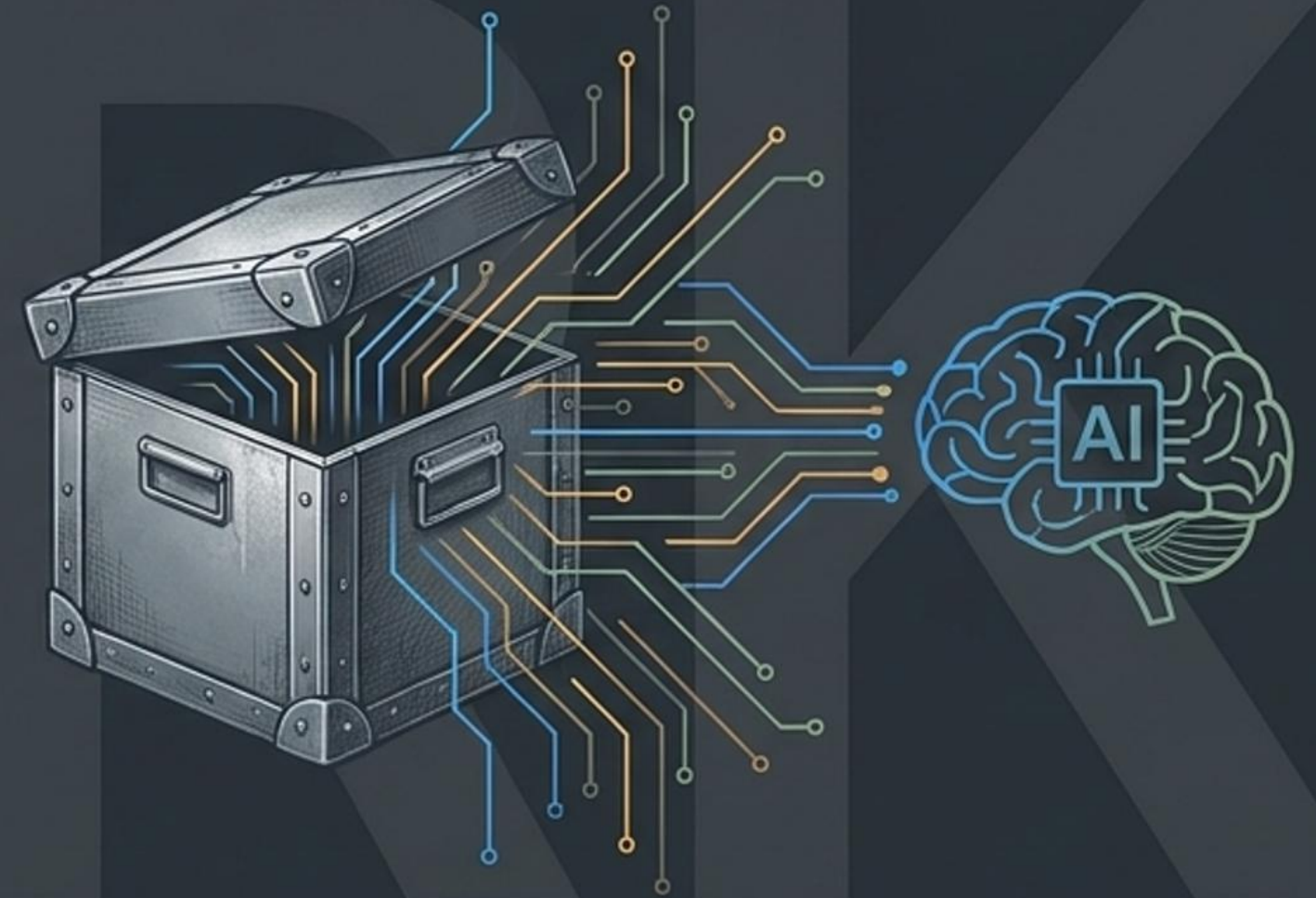


400 1000

Data Awareness: Making Data FAIR for AI



Data that is unclassified, untagged and uncataloged are dark data that becomes increasingly hard to access, accelerate and abstract



AI Data Engine making distributed, diverse and dynamic data findable (catalog), accessible (context), interoperable (content) and reusable (composition)

Data Awareness: a Journey towards FAIR

1 CATALOG

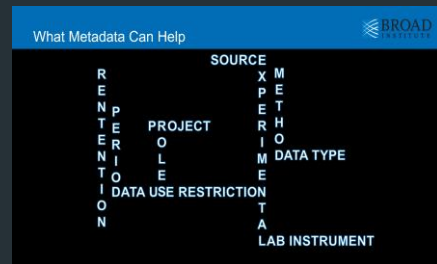
Findable w catalog



metaEngine pilot @ MD Anderson
“finding needle in the haystack”

2 INTEGRATE

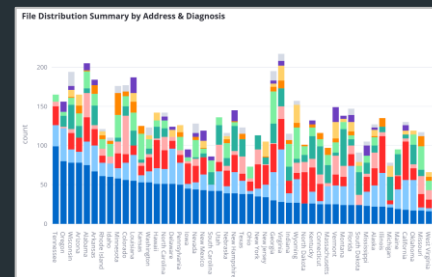
Accessible w context



metaOcean pilot @ Broad Institute
“clean up data swamp”

3 DISCOVER

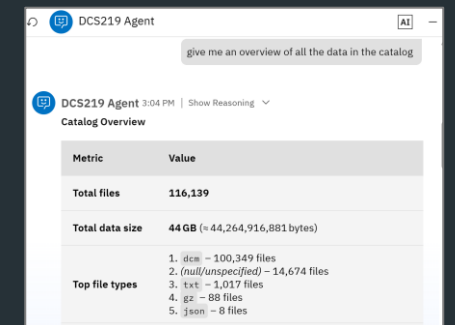
Interoperable w content



RDM + Data Fabric for
Salento at MDA Moonshot

4 ORCHESTRATE

Reusable w composition



AI Chatbot & Avatar for
research data & workflow

2017

2020

2021

2022

2024

2025

2026

Launch of IBM **Spectrum Discover**

Containerized Discover became **Fusion Data Cataloging (FDC)**

IBM launched **Content-aware Storage (CAS)** as AI data service

AIDP & MCP Server for Data Cataloging & Content Aware Storage

give me an overview of all the data in the catalog, including breakdown by 1) file type, 2) size, 3) site, 4), timestamps



Type something...



Q1. List all the imaging records

CAS Demo Case 1: Semantic Search Engine of Knowledgebase

Find all documents related to National Research Platform (NRP)

🕒 Worked for 14.2 Seconds

I have identified 3 matching documents in the repository:

Title	Doc ID	Vector Store	Match Summary
Seungmin_Kim_APAN59_National_Research_Platform_Public.pdf	23071776	gov-demo	This document provides an overview of the NRP's evolution into a distributed cloud supercomputer, detailing GPU growth, network capabilities, and campus integration.
Towards_a_Globally_Distributed_Open_Cyberinfrastructure_2024.pdf	23071774	gov-demo	This document covers the NRP vision and execution strategy, including horizontal and vertical openness, the Open Science Data Federation, and community partnerships.
Status_of_National_Research_Platform_2024.pdf	23071775	gov-demo	This document presents the status of the NRP, highlighting GPU resource expansion, the Open Science Data Federation, and integration with platforms for AI and climate.

CAS Demo Case 2: Multi-modal Insight Engine

Where did "4.5PB distributed data origin across three sites with 100Gbps NVMe caches." come from?

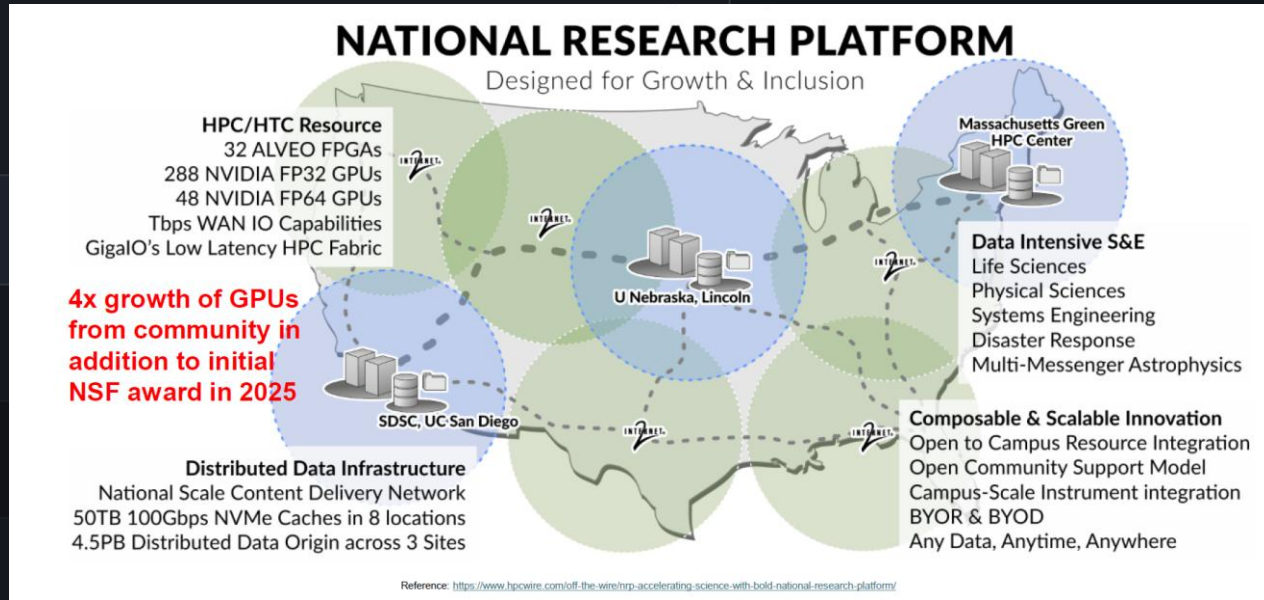
Worked for 13.6 Seconds

The specific data regarding the 4.5PB distributed data origin and 100Gbps NVMe caches originates from documentation describing the National Research Platform (NRP). This platform has evolved from its original vision as a "ScienceDMZ data freeway" into a distributed cloud supercomputer infrastructure.

The infrastructure utilizes a national-scale content delivery network to support high-performance data access across multiple institutional sites. These specifications represent the platform's capacity to handle data-intensive scientific research and large-scale operations across its distributed nodes.

National Research Platform Infrastructure

Feature	Specification
Distributed Data Origin	4.5PB across 3 sites
NVMe Cache Capacity	50TB in 8 locations
Network IO Capability	Tbps WAN IO
Cache Speed	100Gbps



S1

CAS Demo Case 3: Integrative Analysis & Reporting

Create a report highlighting top trends, challenges and opportunities for AI research and education for National Research Platform (NRP) institutions

🕒 Worked for 43.5 Seconds

I have identified the key themes from the provided context to construct this report. Based on the analytical intent, I will synthesize the information into a structured report with sections on trends, challenges, and opportunities.

This report analyzes the key trends, challenges, and opportunities for AI research and education for institutions connected to the National Research Platform (NRP), based on the provided context documents. The NRP is evolving from a data network into a distributed, open-community cyberinfrastructure, with a strong current focus on expanding AI education and building a skilled workforce.

Emerging Trends in AI Research and Education

The context reveals several clear trends in how AI is being integrated into research and education within the NRP ecosystem. These trends point towards a more open, distributed, and education-focused infrastructure.

Trend	Description	Source
Distributed Supercomputing	The NRP is evolving beyond a "ScienceDMZ data freeway" into a distributed cloud supercomputer, enabling composable and scalable innovation across institutions.	S3
Community GPU Growth	There has been a significant 4x growth in community-contributed GPUs, supplementing the initial NSF award and greatly expanding available computational power for AI.	S3
Focused AI Education	There is a strong emphasis on building AI education directly on the NRP, with specific projects like GP-Engine (supporting 7 states) and the CENIC	S1

AI Text Search

Search Type

Select one

- Texts
- Images

Sample Queries

Display Option

Entries to show



Columns to display



Query History

- dogs
- doge
- dogs running in park
- show me photos of dogs running on the beach
- show me photos of dogs playing in the park
- show me photos of dogs playing in the park by the river




Creating a “portable” file-base (filesystem+database)


The screenshot shows a 'File Browser' interface with a dark theme. At the top, there are two tabs: 'File Browser' (active) and 'Selection Review' with a '0' indicator. On the left side, there is a 'SEARCH RECORDS' section with a search input field containing 'f:Filename...'. Below that is a 'SELECT PATIENT(S)' section with a list of four patient IDs, each with an unchecked checkbox: 001075, 001174, 001351, and 008327. The main area displays a directory tree starting with a root folder '/'. Underneath is a folder named 'gpfs', which contains a sub-folder 'fusion111'. Inside 'fusion111' is a folder 'dat-seda', which contains five sub-folders, each with a patient ID label: '001075 Patient: 001075', '001174 Patient: 001174', '001351 Patient: 001351', '008568 Patient: 008327', and '008569 Patient: 008327'. A white mouse cursor is hovering over the '001351' folder.

IBM Global Data Platform for AI

AI Training

HPC Cluster DGX/HGX Node with GPU


AI Inferencing

IBM watsonx **Red Hat AI Enterprise** **NVIDIA AI Enterprise**
 LLMd  Red Hat Dynamo 

Data Awareness

Data Catalog Content-aware MCP Presto / Spark

Data Access

POSIX | S3 | NFS | SMB | HDFS | GPUDirect | CNSA/CSI | Vector

Data Acceleration Tier | KV Cache

ILM | AFM

IBM Global Data Platform

Data Acceleration

Data Abstraction



Commodity Storage HW



Cloud Data Store



Other Vendor Storage HW



Tape Storage